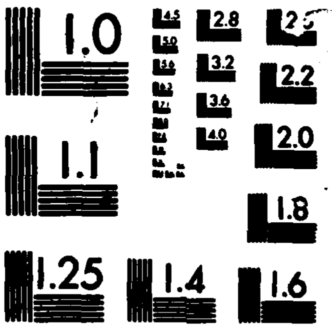
AD-A167 056 HANDLING BIAS FROM INDIVIDUAL DIFFERENCES IN 1/1
BETWEEN-SUBJECT HOLISTIC EXPERIMENTAL DESIGNS(U) ESSEX
CORP ORLANDO FL C W SIMON ET AL. 30 OCT 85 EOTR-86-1
UNCLASSIFIED N61339-81-C-0105 F/G 5/9 NL



MICROCOP' CHART

(12)

ESSEX ORLANDO
TECHNICAL REPORT
EOTR-86-1

HANDLING BIAS FROM INDIVIDUAL DIFFERENCES IN BETWEEN-SUBJECT HOLISTIC EXPERIMENTAL DESIGNS

Charles W. Simon

Daniel P. Westra

FINAL REPORT

30 October 1985

N61339-81-C-0105

Prepared for:

Naval Training Systems Center
Orlando, Florida 32803

Essex Corporation
1040 Woodcock Road
Orlando, Florida 32803

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

AD-A167056

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION/DOWNGRADING SCHEDULE

3. DISTRIBUTION/AVAILABILITY OF REPORT
[illegible] unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)

5. MONITORING ORGANIZATION REPORT NUMBER(S)

6a. NAME OF PERFORMING ORGANIZATION
Essex Corporation

6b. OFFICE SYMBOL *(If applicable)*

7a. NAME OF MONITORING ORGANIZATION
Naval Training Systems Center

6c. ADDRESS *(City, State and ZIP Code)*
1040 Woodcock Road
Orlando, FL 32803

7b. ADDRESS *(City, State and ZIP Code)*
Orlando, FL 32813

8a. NAME OF FUNDING/SPONSORING ORGANIZATION

8b. OFFICE SYMBOL *(If applicable)*

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER

8c. ADDRESS *(City, State and ZIP Code)*

10. SOURCE OF FUNDING NOS.

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT NO.

11. TITLE *(Include Security Classification)* Handling Bias from Individual Differences in Between-Subject Holistic Experimental Designs (Unclassified)

12. PERSONAL AUTHOR(S)
C. W. Simon and D. P. Westra

13a. TYPE OF REPORT
Final

13b. TIME COVERED
FROM ________ TO ________

14. DATE OF REPORT *(Yr., Mo., Day)*
30 October 1985

15. PAGE COUNT
88

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB. GR.

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
Individual differences, subject variability, subject-related bias, subject-factor confounding, holistic experimental designs, (cont'd)

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

The effects of expected biases from individual differences are quantified in this report. Techniques for handling them are discussed, particularly in the context of 2^{k-p} holistic experiments.

Advantages of a holistic approach for equipment design research are presented. In contrast with the traditional few-factors-at-a-time approach, this approach has as its fundamental philosophy the need to investigate "all" of the potentially critical factors in the same experiment. A sequential strategy and bundle of techniques make the approach feasible. To follow this philosophy increases the generalizability of the results and reduces the dangers of misinterpretation when critical factors are held constant. Data collection costs are also reduced over what they would be for equivalent information collected a few factors at a time.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT
UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. □ DTIC USERS □

21. ABSTRACT SECURITY CLASSIFICATION
Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL

22b. TELEPHONE NUMBER *(Include Area Code)*

22c. OFFICE SYMBOL

18. Subject terms (contd)
between-subject designs, covariates, normal-order statistics

19. Abstract (contd)

To make this multifactor approach economical, the experiment is run initially using only one subject per experimental condition. The obvious confounding of the subject and configuration effects in the performance scores creates a concern regarding biased results and the risk of misinterpreting them.

Computer simulations and analyses were performed to better understand the anatomy of the subject-related bias problem and to evaluate ways to reduce it. It is shown how the bias problem is not unique to the holistic approach nor a consequence of using only one subject per cell; instead it is common to all experiments in which different subjects are tested on different experimental conditions. Selecting and assigning subjects at random to experimental conditions does not eliminate the problem; it ensures it. Equations are provided to relate the variability of the subject population to expected biases.

Seven techniques that may be used singly or in combination to reduce the dangers of subject-related bias are described and evaluated. Depending on how much information is available regarding subjects' abilities, the investigator can (with some information): (1) include suspected sources of subject variability as factors in the experimental design; (2) partial out subject effects using a covariate; (3) use more homogeneous subjects; (4) use covariate information probabilistically to reduce interpretation errors; and (with no information): (5) add more subjects per condition; (6) keep the holistic design undersaturated; (7) expand the fractional factorial design. Equations are provided to quantify the effects the above techniques have on the reduction of subject-related bias.

In the appendices, special topics are treated: (A) Computer programs are provided that can be used to determine expected values, their standard deviations, and the proportion of variance accounted for at each rank. (B) The tables for the expected values of order statistic from a normal distribution are given along with support data for experiments containing 31, 63, and 127 experimental effects. (C) The use of "normal order plots" is presented as a means of evaluating the significance of a set of experimental effects when no independent estimate of error is available. (D) The requirements and inadequacies of covariate scores as a means of reducing subject-related bias are discussed. (E) Tables are provided to give the probabilities that a true bias effect will actually be where the covariate bias effects indicate that it is, depending on the number of effects being considered, the validity of the covariate, and the size of the experiment.

TABLE OF CONTENTS

TABLE OF CONTENTS (cont'd)

LIST OF ILLUSTRATIONS

LIST OF TABLES

PREFACE

The authors wish to acknowledge the contributions of: Mr. Brian Nelson of the Essex Corporation, who ran many of the computer programs used in the analyses; Dr. William Dunlap of Tulane University, who provided a number of the computer programs used in the analyses; and Dr. E. Christopher Poulton, Appplied Psychology Unit, Medical Research Council, Cambridge, England, who critically reviewed this report.

SECTION I

INTRODUCTION

Transfer of training experiments were being conducted before the turn of the century. These experiments are intended to answer some form of the question: How well does training on one task facilitate performance on a second task?

Transfer studies have been made in many contexts. For example, can a motor skill acquired by one part of the body be transferred to another part of the body? Will learning one list of words make the learning of a second list easier? Does training on one perceptual-motor task speed the learning of a second perceptual-motor task? Will solving one set of intellectual problems facilitate the solution to another set of intellectual problems? Associated with each of these questions is an interest in the characteristics that create positive and negative transfer.

Much of the early practical work on transfer was done by those evaluating educational methods. With the advent of the second World War, research began to focus on finding better ways to train people to use highly complex equipment associated with the defense effort. Subsequently, transfer experiments were employed to measure and compare the relative effectiveness of alternative training procedures and/or devices on the subsequent performance of operational tasks.

As the cost and complexity of training devices increased -- keeping pace with the operational systems -- it was no longer sufficient to conduct research primarily to evaluate completed systems. Experimental data were needed early in the developmental phase of system design before the final configuration was chosen to determine which simulator features were responsible for the greatest amount of positive transfer as the trainee moved from the simulator to the operational equipment.

To date, this simulator research has not provided the definitive answers expected or required. That simulators as a group are cost-effective tools for flight training is generally accepted today, but determining which factors are responsible for this effectiveness still generates considerable controversy (Caro, 1977; Orlansky, 1982). Typically, one finds meager experimental data supporting either side of the more controversial issues of simulator design. Much of this

TABLE OF CONTENTS (cont'd)

empirical ambiguity can be traced directly to the narrow limits of the experiments and the methodologies employed. Although the purpose of many transfer of training experiments has shifted from postdevelopment evaluation to predevelopment design recommendations, conventional experimental paradigms employed half a century ago continue to be used although they are no longer appropriate.

In this report, the inadequacies of the traditional paradigm will be discussed. A new, more effective transfer of training paradigm for equipment design research used at the Naval Training Equipment Center will be described. Solutions to certain problems of data interpretation that derive from testing a different subject (or subjects) on each experimental condition will be presented.

CONVENTIONAL TRANSFER OF TRAINING DESIGNS

Campbell and Stanley (1963), in their classic book on experimental designs, offer criteria and deficiencies in designs primarily suited for transfer of training research. Although they suggest that more elaborate designs are possible, the paradigms actually employed by behavioral scientists in most transfer of training experiments have traditionally been some minor variation on the design shown in Table 1.

TABLE 1. CLASSIC TRANSFER OF TRAINING PARADIGM

		Pre-test	Training	Post-test
Experimental Group	E	$Ø_1$	X	$Ø_2$
Control Group	E	$Ø_1$		$Ø_2$

Each group contains a number of subjects, presumably drawn from a representative population, the exact number quite frequently being determined by logistic rather than statistical considerations. Depending on the number of groups involved, the number of subjects per group has tended to range between three and ten. Each group performs the sequence depicted above. The E indicates that some effort is made to equate critical differences in subject ability among the groups by assigning subjects in some quasi-random manner, by matching, or by some postexperimental statistical manipulation.

The X shows that the experimental group receives some special training not given to the control group. In some experiments, the investigator may decide to include more than one experimental group, each receiving a different training procedure or configuration. At times, when an old and a new treatment, device, or process are being compared, both groups may receive training appropriate to the condition being studied. The use of the control in that case is often eliminated. In some studies, the multiple experimental configurations can be dimensionalized to form a factorial design, thereby enhancing the information made available. Until recently the total number of experimental conditions (whether dimensionalized or not) has seldom exceeded eight.

The Os in the above design indicate that performance is measured on the criterion (transfer) task both before (i.e., $\emptyset_1$) and after (i.e., $\emptyset_2$) training for the experimental group, and twice (but without the intervening training) for the control. In some training studies, however, particularly when extremely complex tasks are to be learned (e.g., flying an aircraft), pretests are not carried out since it is presumed that the subject cannot perform the task at all without training.

LIMITATIONS OF THE TRADITIONAL PARADIGM

The traditional experimental paradigm is most effective when it is employed to examine *completed* systems. It is most suited for answering questions of the type: Is performance on criterion task, Ø, better than that obtained by the control group after operators have been trained on task X? Will it be cost effective to substitute a new training system for an old one? Does the training effectiveness of an expensive system justify its added costs over a less expensive version? The simplistic nature of the experimental design makes it necessary to assume -- unjustifiably -- that the same answer will hold under a variety of environmental conditions or when critical factors held constant in the experiment take on different values in the operational situation.

The traditional transfer paradigm that compares whole systems is not cost effective when one wishes to examine the transfer effectiveness of simulator components under a variety of operational conditions. In some studies, several components have been varied, but far fewer than the number likely to affect task performance. This limitation is imposed by an inappropriate methodology which makes larger multifactor experiments too costly to consider. Unfortunately, it is from the results of these very limited experiments that investigators and system engineers have drawn inferences regarding the designs of new simulators.

WEAKNESS OF FEW-FACTOR EXPERIMENTS

Because the traditional approach has made it prohibitively expensive to study many factors at a time, most behavioral scientists have been content to perform a series of partially overlapping few-factors-at-a-time experiments. Implicit in such experiments is the assumption that the experimental results and the interpretations of these studies will not be influenced by any of the potentially critical factors held constant.

But this is not necessarily so. The factors held constant do not just disappear. When they are held constant, they must be held at some specific value. Whether selected in a haphazard or purposeful fashion, the value used can markedly affect the experimental results obtained from the factors that were varied and may lead to incorrect generalizations when the results are applied to field situations not falling within the experimental space.

Erroneous conclusions may be drawn from experimental results when:

1. Unrevealed interactions exist between the constant and varied factors.

2. The overall level of task difficulty forces performance into asymtotic limits.

These are illustrated in Figure 1.

Question: Does Level 1 or 2 of Factor A yield the higher performance?

		Factor A (Varied) 1)	Factor A (Varied) 2)
Factor B held constant at	Level 1)	5	10
	or		
	Level 2)	10	5

CASE I: Factors A and B interact

	Factor A (Varied) 1)	Factor A (Varied) 2)
1)	5	10
or		
2)	5	5

Case II: Factor B Affects Overall Difficulty Level

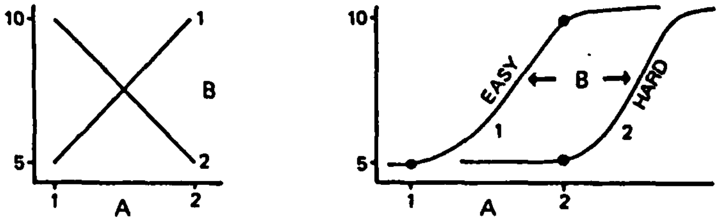


Figure 1. Effects of factors held constant.

Note how the conclusion regarding Factor A would differ depending on whether the level of Factor B were fixed at Level 1 or Level 2. In the first example, the investigator would have no knowledge of the interaction between A and B and would draw conflicting conclusions depending on the level at which Factor B had been held constant. While it might be argued that an investigator must recognize that it is dangerous to extrapolate the results beyond the experimental space, in practice, the "situation specific" nature of human behavior is usually forgotten, and results are frequently generalized to situations to which they do not apply. (Just contemplate the information applied to the training of experienced pilots flying high-performance jet aircraft that came from studies using a simplified version of a Link trainer flown by inexperienced pilots.)

In the second example, no real interaction exists. Still the value at which Factor B is held constant can influence the size of Factor A's effect. The value of Factor B will affect the level of overall task difficulty. If the task becomes too easy or too difficult, then performance can hit a ceiling or floor (as in Case II, Figure 1) where potential differences between two conditions of Factor A are no longer apparent.

When we consider the fact that there are usually not one but a great many factors held constant, often without careful planning on the part of the investigator, the potential for misinterpreting the data increases unless prudent procedures are followed. Furthermore, unless the values for the factors held constant in the experiment correspond to those encountered in the real world, attempts to predict real-world performance will be biased by the differences.

THE NEED FOR MULTIFACTOR TRANSFER EXPERIMENTS

If we wish to obtain valid and generalizable data on the transfer effectiveness of a large number of system components, a truly multifactor experiment is needed with ranges of values that cover the space of concern in the real world. Some form of factorial study is needed to detect interactions and to measure the components in combinations that will reflect those interactions. One must attempt to include "all" of the potentially critical factors in the same experiment to make valid generalizations to the operational situations.

In the early 1950's Williams and Adelson (1954) were asked by the Air Force to examine the requirements for fidelity in a flight simulator used for pilot training. As part of their analysis, they describe a transfer of training experiment to determine the relative savings for various simulator configurations involving 34 simulator characteristics. In their report, the investigators explained the rational process they went through in planning such an experiment. They wished to vary each of the 34 simulator characteristics over five values to map the function interrelating these characteristics. However, noting that a factorial experimental plan would involve 5.8×10^{23} combinations, they stated that "it is manifestly impossible to conduct a transfer of training experiment for each [of the conditions of the full factorial], since there would be neither sufficient time nor enough subjects to complete the project" (p.8). They then explored the possibility of looking at each factor individually and testing 20 pilots at each level. Since this also proved to be too expensive to be practical, they considered limiting their investigation to "only the important" factors. To this, they concluded: "But here a dead end is encountered for there is no a priori way to decide which are the important characteristics that should be studied" (p.8). They noted that they had already selected the important factors and had kept the number at a minimum. Furthermore, they pointed out that even if such a study were possible, it would be limited to only a few flight tasks and a specific aircraft simulator. For these reasons, they wrote, "one hesitates to recommend purchase of a variable characteristic simulator for the purpose of studying fidelity of simulation" (p.9). No experiment was ever conducted.

This case exemplifies the dilemma faced by psychologists concerned with an empirical determination of training simulator requirements. On the one hand, dimensionalizing transfer of training experiments adds markedly to the validity of the information obtained. On the other hand, for large multifactor experiments, the size of the data collection effort using the traditional experimental paradigm becomes too prohibitive to be practical.

A HOLISTIC EXPERIMENTAL PARADIGM

A holistic paradigm for conducting large multifactor experiments has been described in several reports by Simon (1977b, 1979). It enables a great many factors, i.e, 10 or more, to be investigated in an integrated, sequential data collection effort at far less cost than would be achieved when studying even three or four factors at a time using the traditional approach. This paradigm has already been employed in experiments at the Naval Training Equipment Center in which immediate simulator performance by experienced operators was the primary criterion (Westra, Simon, Collyer, and Chambers, 1981; Westra, 1982).

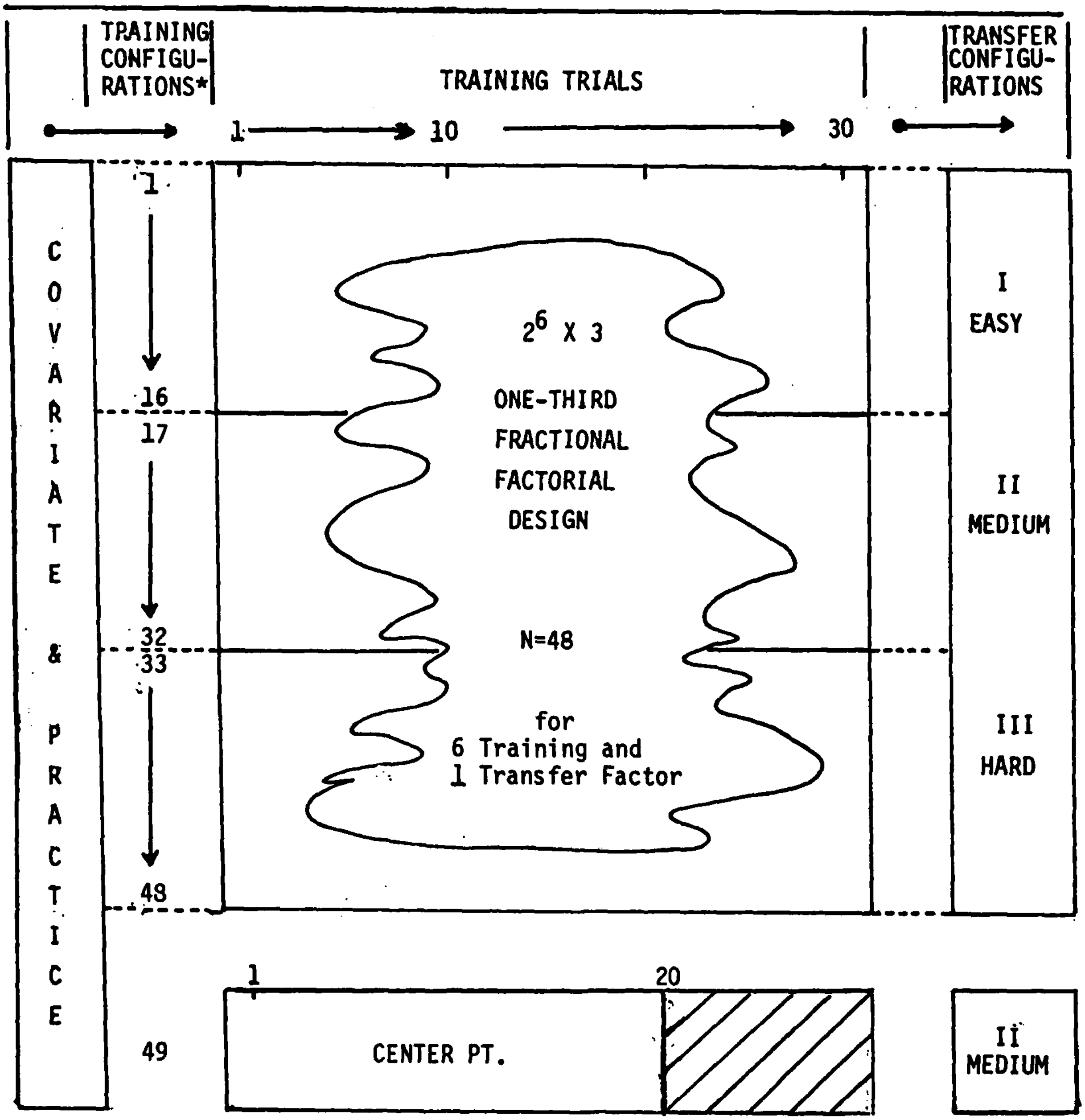
The same philosophy, strategy, and many of the techniques of the holistic paradigm can be applied to transfer of training experiments using inexperienced operators. The term "holistic" has been used by Simon to identify a particular paradigm composed of a philosophy, strategy, and bundle of techniques for conducting large scale, multifactor, controlled experiments economically. It is not an experimental design but a practical methodology for the design of experiments. It provides a pragmatic, empirical description of performance on a designated task throughout the multifactor (and multivariate) space, properly protected against such sources of bias as lack-of-fit of the experimental model along with time and subject bias effects. The holistic approach has the same advantages for transfer experiments as it does for immediate-performance experiments insofar as quantity, quality, and costs of information are concerned. However, an additional strain is put on the size of any experimental effort in transfer experiments over that experienced with the immediate-performance studies. This is due to the additional time required to train subjects on different configurations, the introduction of the transfer task, and the fact that a different subject is being tested on each training condition. Still, if the problem is important enough, the effort is justified.

APPLYING THE HOLISTIC PARADIGM TO A TRANSFER EXPERIMENT

Simon and Roscoe (1981) conducted a laboratory experiment to demonstrate the advantages of the holistic approach to transfer of training experiments. A quasi-transfer of training study was carried out in which both training and transfer were measured on different configurations in the same simulator. The task was a horizontal tracking test, and a total of 80 college students were used as subjects in the experiment. The investigators trained a different subject on each of 49 different simulator configurations in an experimental design that permitted both training and transfer performance to be mapped over a seven-dimensional space.

The seven factors were: vehicle control order, display lag, tracking mode (percent pursuit versus compensatory), prediction time, control gain, training trials, and difficulty -- at three levels -- of the criterion (transfer) task. Three control groups, one for each criterion task, were tested. The experimental design is shown conceptually in Figure 2. The cost-to-information ratio exhibited by this data collection plan is a marked improvement over anything possible using traditional transfer of training designs.

Westra (1982) employed a similar plan in an in-simulator experiment to study the transfer effectiveness of six simulator design features for training two types of pilots (with no prior carrier-landing experience) to land on a simulated aircraft carrier. The six features were field of view, motion, scene detail, turbulence, glideslope rate cuing, and approach type. An experimental design made up of 32 training configurations was used to evaluate the transfer effectiveness of all critical combinations of the seven factors (including pilot types) and their two-factor interactions. A single criterion task -- a simulator configuration judged to have the highest fidelity to a



*[One subject per configuration except #49 = 8 subjects]

Figure 2. Transfer of training holistic design.

real aircraft carrier-landing task -- was used. A different subject was trained on each configuration.

Using the Simon and Roscoe experiment as the primary example, the items listed below illustrate the power and some of the advantages of using a holistic approach in transfer of training experiments:

1. Informative: The data from the 48 subjects/training configurations were sufficient to estimate all main and two-factor interaction effects for six training factors and three transfer configurations for training and transfer performance. By adding eight more subjects at the center point of the experiment, the probable presence of some higher-order effects could be tested.

2. Precise. By taking advantage of the "hidden replication" in the factorial-type design, all estimates of an equipment factor's main performance were based on 24 performance scores, and all estimates of main and two-factor interaction effects were based on the differences between 24 pairs of values. Estimates of transfer means were each based on 16 performance scores.

3. Generalizable. An analysis of the data produced an equation that provides an estimate of the transfer effectiveness of any training configuration within the boundaries of the experimental space -- whether empirically studied or not -- on the performance on three criterion (transfer) configurations.

4. Economical. For screening purposes, the training configurations that yield the greatest amount of transfer to any of the criterion tasks can be identified (within the confidence limits of the experiment) without ever testing a control group. This makes an already cost-effective design still more economical. This is so because an adjustment for a control group merely removes a constant from all the transfer scores. Thus, relative performance is known whether the control data are involved or not. (Control data need be collected only to estimate savings, a step that should be delayed until the system evaluation phase begins. At that time, the few training configurations tentatively identified during the screening phase would be examined more stringently.)

SECTION II

THE PROBLEM

In multifactor transfer experiments involving 32, 64, or more conditions, the need to train different subjects on each experimental condition can frequently put a strain on subject availability, however economical a design may be. Furthermore, with the extended training sessions and additional transfer trials, the overall time to do a transfer of training experiment is markedly increased over that for a simpler, immediate-performance study using the same basic experimental design. For these reasons, to avoid having to reduce the number of potentially critical factors being investigated, the experimenter initially will use a single subject per training condition. Then only after examining his data will he exercise the option to add more subjects per condition, if necessary.

Two arguments against using only one subject per condition are frequently raised, although neither is crucial to our experimental goals in the early stages of the research program. The first is that one subject per condition does not provide a very reliable estimate of performance on a single training configuration. The second is that with one subject per cell, no within-cell variance estimate is available as an error term for testing the significance of the effect.

The first argument misses the point of the screening experiment (Simon, 1977a). At this stage of an investigation, the experimenter seeks empirical data which can help him identify the few truly critical factors out of a great many candidates for the task at hand. He is only casually concerned with precise estimates of performance on individual configurations, although in fact the estimate from a properly designed screening study will be quite precise. Each main- and two-factor interaction effect in these two-level designs is the mean difference of N/2 values, with N being the total number of observations in the experiment. An estimate of the error variance in these experiments can also be expected to be based on at least N/2 degrees of freedom. For holistic experiments of 64 conditions or greater, the reliability of the estimates should compare quite favorably with those obtained in studies of typical human performance.

The coefficients of the factors identified as critical can be written in equation form and used to estimate performance on any configuration within the total experimental space *whether* actually studied or not. While these estimates will be relatively reliable, one might better limit their use to the primary purpose of the screening design -- determine the critical factors and localize a much more limited space within which a more stringent investigation would focus. Absolute performance values are better obtained in the evaluation stage of the investigation, where the configurations selected from the screening data are studied in depth, preferably in an operational environment.

The second argument, that using one subject per condition provides no measure of within-cell variability (to be used as an estimate of the error variance), is true in fact but false in implication. There are a number of alternative techniques available for estimating error variance that do not require a total replication of the design. Center point replication and other forms of partial replication techniques have been discussed (Simon, 1973). The normal-order plots (Daniel, 1959, 1976; Simon, 1977a) used to estimate the statistical significance of an experimental effect also provides a reasonable measure of the error variance when the larger ($C > 64$) experimental designs are involved even for unreplicated data (see Appendix C).

While neither of the above objections represents a serious cause for concern, a third, more surreptitious problem relating to subjects does exist in the transfer of training paradigm described earlier and could assume unacceptable proportions if not properly handled. This is the bias that inevitably occurs whenever a different subject (or subjects) is randomly assigned to each of the experimental conditions.

BIAS FROM INDIVIDUAL DIFFERENCES

That individual differences can complicate the conduct and interpretation of human performance research is well recognized. Although differences among subjects are merely one of many potential sources of variance and bias in experiments, the unique characteristics of human beings present special problems that distinguish behavioral science research from that of the physical sciences. This is not because the paradigm for experimental designs, *per se*, will be different; it is just that in general "human factors" are less identifiable, less measurable, less manageable, and less understood than those ordinarily investigated in physical science experiments.

When individual differences are not the primary focus of an experimental investigation, the investigator may try to control them in various ways:

- by using the same subjects across all (or blocks of) conditions,

- by assigning subjects so as to equate abilities across conditions,

- by assigning subjects at random to the experimental conditions,

- by removing subject differences statistically after the experiment has been completed.

In transfer of training experiments, certain limitations occur. It is ordinarily not feasible to train a subject on more than one configuration or training procedure. Frequently, the specific ability of the subject who is about to be trained to perform the task for the first time is neither known nor measurable, making equating essentially impossible. The use of one subject per cell to economize when a great many conditions are being studied makes it still more difficult to equate among the cells for individual differences. The use of covariates will be discussed in another section, while the inadequacy of randomization is discussed below.

BIAS VS VARIABLE ERROR. When different subjects are distributed randomly among the experimental conditions, the effects of individual differences can both bias the experimental effects and contribute materially to the size of the error variance term. Traditionally, more concern has been given to the variable error component created by individual differences, particularly as it affects the power of the tests of statistical significance. In this report, the effects of <u>bias</u> error will be the point of focus.

To say that individual differences bias the effects means that the differences in performance between different levels of an experimental factor have been confounded with mean differences in ability of the groups assigned at random to the factor levels. While this is nothing new in performance experiments, the holistic approach exposes this bias more vividly. In the traditional approach the problem tends to remain submerged and muted by the excessive redundancy of the data collection. The luxury of considerable replication is not a viable choice with the holistic paradigm.

Just how the effects of individual differences distribute themselves in a single data collection sample is shown in the following familiar equation, representing the F-ratio employed in a test of statistical significance:

$$F = \frac{\text{B/Cell Var. (i.e., Factor Var. + Subj. Var. b/cells + Error Var.)}}{\text{W/Cell Var. (i.e., Subj. Var. w/cells + Error Var.)}} \qquad \text{(Eq. 1)}$$

All sources of variance within each set of brackets are confounded; that is, they cannot be independently measured nor isolated. Subject variability is confounded with factor variability in the numerator and inflates the true error variance in the denominator. We must not be tempted to forget that subject variance is not a source of random error or unknown variance; it is a known source of variance that cannot be isolated in this application.

In Table 2, a fictitious example is given to illustrate what is happening in Equation 1. For this example, the eight experimental conditions of a 2^3 factorial design will represent different equipment configurations. The three tables show the contributions to each cell made by the subjects (2.1), the equipment configurations (2.2), and the two sources combined (2.3). Beneath each table, the total between and within cell variances for that table are presented.

The data shown in the tables were obtained as follows: Sixteen performance values were chosen to approximate that which would be expected had 16 subjects of different abilities been selected at random from a normal population with a mean of zero and a standard deviation of one.* These numbers were assigned at random to the eight experimental conditions, two to a cell (Table 2.1). The sum of both subjects' scores within a cell is shown. The breakdown of variances between and within cells is shown below each table. For this sample, the total subject variance was approximately .92 (close to but not exactly the population variance of one), and total subject variability was distributed unevenly between and within cells.

Theoretically, when subjects selected randomly from a normal population are assigned at random to the experimental conditions, the expected between- and within-cell subject variances are equal to each other and the population variance. Because the sum of within- and between-subject sums of squares equal 100% of the total subject sums of squares, any differences between these sources behave reciprocally. Thus, if subject bias gets larger, error variance due to subjects gets smaller, and vice versa, and significance tests will be distorted either way. Assigning an error probability of .05 to the F-test of statistical significance does not escape the problem as some would like to believe. It helps us little to know that our chance of drawing erroneous (Type 1) conclusions is five in 100,

*Throughout this report, the notation: N[0,1] is used to describe a population with a mean of zero and a variance (and standard deviation) of one. The same notational form but with different numerical values will be used for normal populations with a different mean and variance.

TABLE 2. THE ANATOMY OF SUBJECT-FACTOR CONFOUNDING (Fictitious data)

Exptl. Design:		(1)	a	b	ab	c	ac	bc	abc
Cell:		1	2	3	4	5	6	7	8
2.1 SUBJECT ABILITY (N[0,1])									
Score	a.	.76	-.99	-.08	1.77	.08	.57	-.23	-1.28
	b.	1.28	.23	.40	-.76	-.57	-1.77	-.40	.99
Sum/Cell:		2.04	-.76	.32	1.01	-.49	-1.20	-.63	-.29

VARIANCE ANALYSIS-- TOTAL: .916; B/Cells: .573; W/Cells: 1.22

2.2 EQUIPMENT CONTRIBUTION

Score	a.	.23	.86	.36	.14	.75	.92	.17	.08
	b.	.23	.86	.36	.14	.75	.92	.17	.08
Sum/Cell:		.46	1.72	.72	.28	1.50	1.84	.34	.16

VARIANCE ANALYSIS-- Total: .113 B/Cell: .243 W/Cell: 0

2.3 CONFOUNDED PERFORMANCE

Score	a.	.99	-.13	.28	1.91	.83	1.49	-.06	-1.20
	b.	1.51	1.09	.76	-.62	.18	-.85	-.23	1.07
Sum/Cells:		2.50	.96	1.04	1.29	1.01	.64	-.29	-.13

VARIANCE ANALYSIS-- Total: .825 B/Cell: .378 W/Cell: 1.22

2.4 EFFECTS FROM ANOVAs OF SUBJECTS, CONFIGURATIONS, AND COMBINED SCORES

Source	Subjects	+	Configurations	=	Combined
Mean	.00		.88		.88
A	-.62		.24		-.38
B	.20		-1.00		-.80
AB	1.13		-.55		.58
C	-1.30		.16		-1.14
AC	.43		-.16		.27
BC	.18		-.41		-.23
ABC	-.61		.29		-.32

since we cannot know where they are in practice. If one is trying to solve real-world problems of import, then we are on dangerous ground by putting our faith in a statistic after the fact rather than anticipating and compensating methodologically for the problem beforehand.

As has been frequently noted, by these standards, and taking journal rejection rates into consideration, more than 5% of the experiments published by psychologists are expected to arrive at the wrong conclusions. If one stops and ponders what this really means for serious investigations of real-world problems, one cannot be content. Even if one had valid and reliable knowledge of subjects' abilities to use to assign subjects to the cells, and thus be able to minimize the biases or minimize the error variance as we wish, we still could not do both in the same experiment. Consequently, however we may assign our subjects to cells, the problem of bias error can be serious.

In Table 2.2, the contribution of each equipment configuration is shown. While only one configuration is represented in each cell, its value is shown twice to parallel the subject values with which it will be combined. From this table, it is obvious why configurations only contribute to the betweeen cell variance; there is NO variance within cells. (Note: To keep the discussion simple, no error variance is included in this example.)

In Table 2.3, the subject and configuration contributions are combined. This table of confounded performance scores constitutes the information that would be obtained when the experimental data are collected.

ANOVA. When we identify the eight cells as experimental conditions of a 2^3 factorial design -- they are named at the top of Table 2 -- and perform an analysis of variance, we obtain the seven effects shown in Table 2.4 for the data in Tables 2.1, 2.2, and 2.3. Note how the confounded effects are the arithmetic sum of the effects due to subjects and to configurations individually. Note also the effect of the confounding. For example, in Table 2.4, the effect of Factor C is the smallest of the unbiased equipment configuration estimates, but when biased by the subject effects, the investigator would perceive it to be the largest effect (even with a different sign). Other distortions due to the confounding are also apparent. The results reveal what was expressed in Equation 1; namely, that the within-cell variance is primarily composed of the within-cell subject variance while the between-cell variance is a combination of factor and subject variance.

This example illustrates two important generalizable points:

Point 1. Randomization does not eliminate this type of bias, but in fact ensures it. (This will be quantified in the next section.)

Point 2. Using more than one subject per cell does not eliminate this type of bias.

When we use only one subject per cell -- a decision usually made because of the need for economy -- we eliminate the denominator of the equation, leaving:

Between Cell Var. (i.e., Factor Var. + Subj. Var. b/Cells + Error Var.) (Eq. 2)

As previously noted, no conventional F-test can be carried out since only the numerator remains. The data from the different cells treated as a 2^3 factorial design can be analyzed, however, and any bias due to subject variability will still be present. Bias, not the error variance (or lack of same), is therefore our primary concern. Neither randomization nor replication provides a complete solution to this problem.

Figure 3 graphically summarizes the process described in this section, showing how subject variability is carried through the steps: population to sample to calculated effects.

ESTIMATING THE SIZE OF THE BIAS FROM SUBJECT-FACTOR CONFOUNDING

Experimenters do not need to be told that subject bias exists. They need to know where it exists, how much exists, when it is likely to be disruptive, and what to do about it. Certainly in the fictitious example given above it had a serious effect.

Computer analyses, including Monte Carlo studies, were employed to obtain some general answers to these and other questions regarding subject-related bias*. A fictitious normal population of subject abilities having a mean of zero and a variance of one was created and sampled at random by the computer. N subject ability scores were sampled from the defined population and assigned at randon to the C conditions of a 2^{k-p} factorial design of corresponding size, initially with one subject per condition. These numbers were treated as data and subjected to an analysis of variance from which (C-1) = K independent main and interaction effects could be estimated.

*Information regarding the computer programs used to prepare the material in this paper may be found in Appendix A.

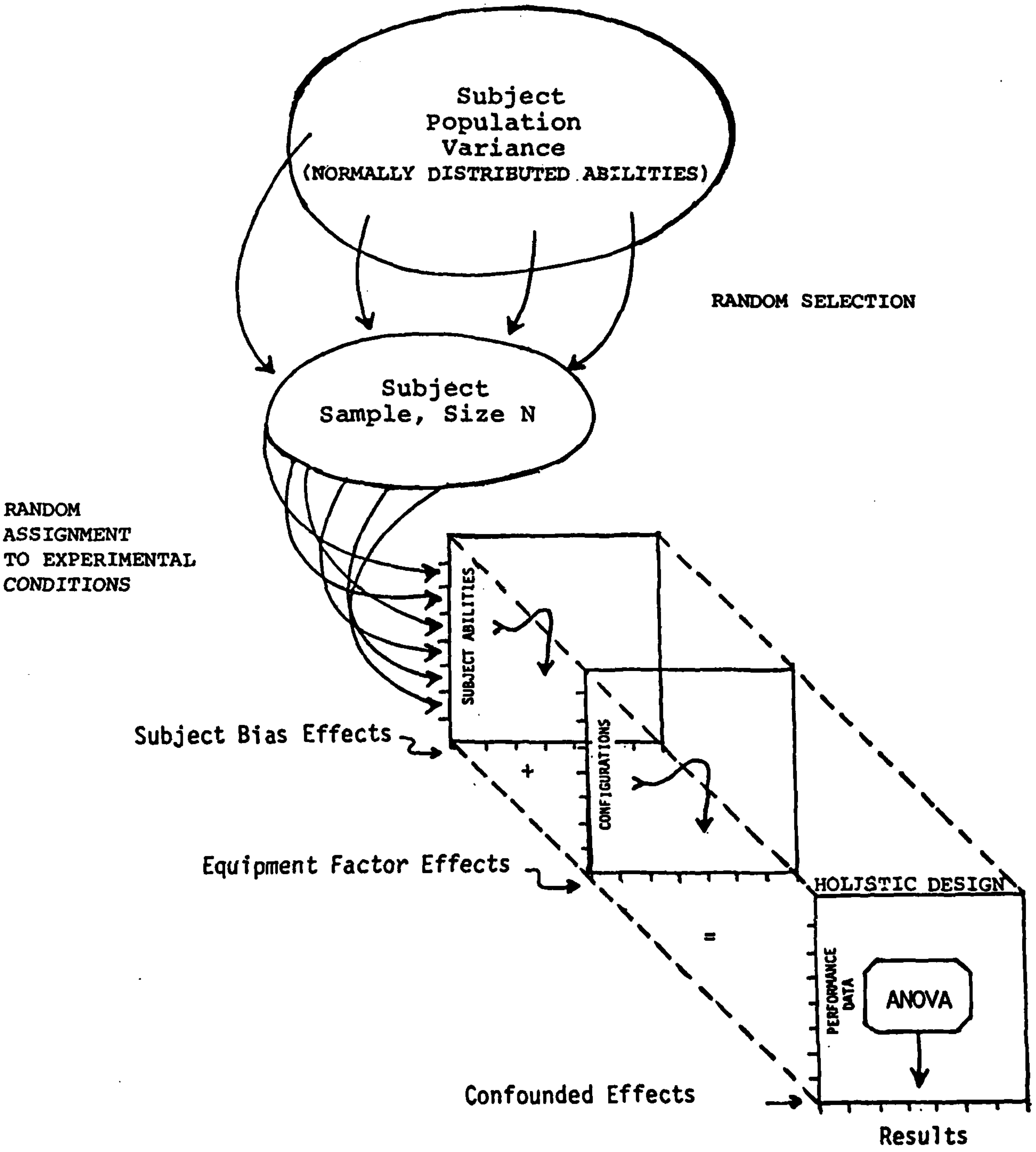


Figure 3. How individual differences confound with equipment configurations to bias equipment factor effects.

These subject effects, i.e., the mean difference between two levels of subject abilities for each effect, were ranked in magnitude with rank one being the largest positive effect and rank K being the largest negative effect that was obtained. Theoretically, these values are symmetrical around zero with opposite signs for each half.

The above process was repeated ten thousand times and the mean of the subject effect (i.e., expected bias), the standard deviation of the bias (s-bias), and the proportion-of-total-variance-accounted-for were obtained for each of the K rank positions. The results of this for N = 16, K = 15 are shown in Table 3.

TABLE 3. ESTIMATED SUBJECT BIAS EFFECTS FOR K = 15, N = 16 (10,000 TESTS, RANDOM SELECTION, AND ASSIGNMENT FROM AN N[0,1] SUBJECT POPULATION)

A Rank by Size	B Expected Size	C Stand. Dev. of Bias	D Prop. Total Var.	E z-score
1	.87	.27	.237	1.74
2	.62	.21	.120	1.24
3	.47	.19	.070	0.94
4	.36	.17	.041	0.72
5	.26	.17	.021	0.52
6	.17	.16	.009	0.34
7	.08	.16	.002	0.16
8	.00	.16	.000	.00
9	-.08	.16	.002	-0.16
10	-.17	.16	.009	-0.34
11	-.26	.17	.021	-0.52
12	-.36	.17	.041	-0.72
13	-.48	.19	.070	-0.96
14	-.62	.21	.120	-1.24
15	-.87	.27	.237	-1.74

The mean of the 10,000 mean differences at each rank is an estimate of the expected bias due to subjects from the N[0,1] population for that rank. The standard deviation of the expected bias at each rank shows how much the sample biases may vary in size around the expected value. The proportion of variance provides a measure of the relative contribution the bias at each rank makes to the total variance.

When subjects from the normal population are assigned at random to the N experimental conditions, the K biases determined by the above analysis will be confounded in some manner with the

K experimental effects. Any experimental effect may be confounded (biased) by a positive or a negative subject effect which, in turn, may be relatively large or small.

Thus we are able at this point to estimate the biases that will occur when subjects are chosen and assigned at random from a normally distributed population (N[Ø,1]), but we have no way of knowing in a real experiment with which of the K experimental effects each bias will be confounded, or for that matter, what size the single sample bias will actually be.

Expected biases can be expressed in standardized form (z-scores). These can be obtained in the conventional way by subtracting the average bias (which in this example was zero) from an expected bias (i.e., mean difference) at any rank and dividing the result by the expected standard error of the subject mean difference. This standard error of the mean differences is obtained by multiplying the population standard deviation (in this case equal to one) by the square root of the quantity four over N, where N is the total number of independent observations in the experiment. The equation for this is:

$$\begin{array}{l}\text{Expected z-score}\\ \text{of bias effect}\\ \text{for rank i of}\\ \text{K cases}\end{array} = \begin{array}{l}\text{Expected bias for}\\ \text{rank i of K cases}\\ \text{from a population}\\ \text{N(0,1)}\end{array} \div \left[\begin{array}{l}\text{Population}\\ \text{Standard}\\ \text{Deviation}\end{array} \cdot \sqrt{\frac{4}{N}}\right] \qquad \text{(Eq. 3)}$$

where K cases are the number of independent effects being seriously examined in a 2^{k-p} experiment, one subject per condition.

In practice, the subject variance and standard deviation will probably not be equal to one, as in the fictitious population. To complete the equations, the investigator must be able to estimate the population standard deviation, which is usually done using data from other similar experiments or is estimated from empirical data obtained from the subjects in the ongoing study.

The z-score values in Table 3 are the standardized scores of the expected biases. Such z-values for various "N"s are published in tables of "expected values for order statistics from a normal distribution" (Beyers, 1961; Harter, 1961; Owen, 1962). If no table for a particular "N" can be found, the computer programs provided or referenced in Appendix A can be used to generate expected values. Tables of expected values for "N" = 31, 63, and 127 along with support data are published in Appendix B.

The expected scores found in the published tables of order statistics for normal distributions can be converted into the expected bias values for any particular 2^{k-p} experiment using a rearrangement of Equation 3:

$$\begin{array}{l}\text{Expected bias}\\ \text{for rank i of}\\ \text{K cases}\end{array} = \begin{array}{l}\text{Expected z-score}\\ \text{for rank i of K}\\ \text{cases from popu-}\\ \text{lation } \underline{N}(0,1)\end{array} \cdot \left[\begin{array}{l}\text{Population}\\ \text{Standard}\\ \text{Deviation}\end{array} \cdot \sqrt{\frac{4}{N}}\right] \qquad \text{(Eq. 4)}$$

Similarly:

$$\begin{array}{l}\text{Standard devia-}\\ \text{tion of bias for}\\ \text{rank i of K cases}\end{array} = \begin{array}{l}\text{Standard deviation}\\ \text{of the z-score for}\\ \text{rank i of K cases}\end{array} \cdot \left[\begin{array}{l}\text{Population}\\ \text{Standard}\\ \text{Deviation}\end{array} \cdot \sqrt{\frac{4}{N}}\right] \qquad \text{(Eq. 5)}$$

The proportion of variance for each rank is obtained as follows:

$$\begin{array}{l}\text{Proportion of}\\ \text{variance for}\\ \text{rank i of K}\\ \text{cases}\end{array} = \frac{(\text{z-score at rank i})^2}{\text{Sum of (z-scores)}^2} \qquad \text{(Eq. 6)}$$

Proportions at several ranks may be accumulated to determine the expected proportion of variance accounted for purely by chance by combinations of the largest effects (either positive or negative or both).

Let us illustrate briefly how the data in Table 3 might be used. For example, when 16 subjects are drawn at random from a population in which subject abilities are distributed $\underline{N}[0,1]$ and assigned at random among 16 experimental conditions, i.e., one subject per cell, the K=15 estimable effects from an analysis of variance will also be distributed randomly and the largest expected positive bias will be 0.87, which is in units of subject population standard deviation. If we did not expect many real effects to be larger than one, for example, then we must do something about the bias since if a +1 effect happened to be confounded with a ∓.87 subject bias, they would essentially neutralize one another and a real effect would not be detected. Or, with a standard deviation of the largest expected bias (rank one) equal to 0.27, there is a .16 chance that the obtained actual bias might be greater than 1.14 which could cause a trivial effect to appear quite important.

If for whatever reason an investigator does not intend to examine all 15 estimable effects, then he would look at a table of normal order statistics for the "N" (in the table) equal to the number of effects (K) that he does intend to examine. The z-values in that table would then be converted into the appropriate bias values, as shown in Equation 4, using the original N of 16 subjects.

For comparison purposes, the four largest positive expected biases (ranks 1,2,3,4) and z-values for K=31, 63, and 127 are shown in Table 4. The largest corresponding expected negative biases are of the same magnitude. The standard deviation of each z-value (s of z) along with the proportion of the variance (Prop.) accounted for by each, and the corresponding largest effects in a 2^{k-p} experiment with one subject per cell, are also given. The complete tables of these expected values are given in Appendix B, the only known published source of tables for K=127. The bias values shown in Table 4 will not be found in Appendix B, but can be calculated using Equation 4. Note in Table 4 that while the larger effects from larger samples are further from the mean (i.e., larger z-values), the biases actually get smaller.

TABLE 4. EXPECTED VALUES AT RANKS 1 THROUGH 4 OF ORDER STATISTICS FROM A NORMAL DISTRIBUTION FOR K=31, 63, and 127

	Rank	z-value	s of z	Prop.	Bias*
K=31					
	1	2.056	.494	.148	.727
	2	1.632	.369	.093	.577
	3	1.383	.319	.067	.489
	4	1.198	.292	.050	.424
K=63					
	1	2.338	.452	.091	.584
	2	1.956	.330	.063	.489
	3	1.739	.281	.050	.435
	4	1.582	.254	.041	.396
K=127					
	1	2.592	.419	.054	.458
	2	2.242	.300	.040	.396
	3	2.048	.253	.034	.362
	4	1.909	.225	.029	.337

* These bias values are specific to two-level experiments with one subject per condition, where the numbers of conditions are 32, 64, and 128, respectively. Complete tables for the other values of normal order statistics for these Ks can be found in Appendix B.

In the aforegoing discussions, we have shown that:

Point 3: When different subjects of different abilities are selected and assigned at random to the various conditions of the 2^{k-p} experiment, the maximum amount of expected bias from subject variability depends on the subject population variance, the number of estimated effects, and the size of the sample.

Point 4: Subject-related bias will occur to some degree in all experiments in which different subjects are assigned at random to the cells.

NON-NORMAL DISTRIBUTIONS. Throughout this report, the assumption has been made that subject populations are normally distributed. In practice we may encounter distributions which depart considerably from the normal. Fortunately, the central limit theorem (Hays, 1963) is operating here. Even with subject population distributions that are considerably skewed or otherwise nonbell-shaped, expected biases will not be seriously altered. Regardless of the abnormalities that may exist in the parent population, the distribution of the means of samples drawn from that population will tend toward normality. (Once again, this reminds us that the tables in this report are means -- what is expected -- although any single experiment may not reflect that value exactly. That is why the standard deviations are given in the tables of expected values.)

Two of the more common non-normal distributions are those that are skewed (B1), i.e., being nonsymmetrical, yielding more measures at one end than another; or kurtotic (B2), i.e., being symmetrical but being relatively peaked (leptukurtic) or flat (platykurtic) in the neighborhood of the mode. For normal distributions, B1=0 and B2=3. When B1 deviates from 0, the data is skewed. For B2 >3, the data is peaked and <3, the data is flat.

Using a Monte Carlo simulation with 10,000 runs, the bias values for a K=31, N=32 experiment were obtained for distributions with different amounts of skewness and kurtosis. The results of these efforts; that is, the eight largest ranks, are reported in Table 5 along with the values for the normal distribution.

The skewed distributions were created by raising the discrete values from a table of order statistics to the second (Skewed 2) and fourth (Skewed 4) powers. The new numbers were then scaled so that the variance would be equal to one and then used to obtain the estimated mean values shown in Table 5. In the skewed distributions, a leptokurtic element is also present.

The kurtotic distributions were created as follows: The leptokurtic (peaked) distribution was obtained by cubing the discrete values from a table of order statistics and adjusting the new numbers so that their variance equalled one. From these, the expected mean values for a peaked distribution shown in Table 5 were obtained. This data is symmetrical, though peaked. The platykurtic data was obtained by taking the cube root of the discrete values on the table of normal order statistics and adjusting them to have a variance of one. From these, the expected mean values for a flattened distribution shown in Table 5 were obtained.

Overall, since some extreme cases of non-normal distributions were used for our examples, Table 5 supports the central limit theorem by showing no serious deviations in the expected mean values for any shaped distribution.

It should be noted that subject outliers that might distort distributions could occur with serious consequences. These subjects, however, could not be considered part of a well-defined distribution and should be dealt with directly. An investigator should not let the robustness for non-normality, as expressed in the central limit theorem, decrease his vigilance in detecting true outliers.

SUBJECT BIAS: HOLISTIC VERSUS FEW-FACTOR EXPERIMENTS

If, in fact, all experiments with different subjects per cell bias the experimental results to some degree regardless of whether one or more subjects are tested on each condition, why should we suddenly become concerned with this matter when conducting holistic experiments? Why wasn't there equal concern when the more conventional few-factor experiments were being conducted?

TABLE 5. EFFECTS OF NON-NORMALITY ON THE LARGER BIAS EFFECTS (K=31, N=32) (All distributions are scaled to have a variance of one.)

	Normal	Skewed 2	Skewed 4	Peaked	Flattened
[B1,B2]	[0,3]	[2.8,5.1]	[7.6,9.8]	[0,8.4]	[0,1.3]
Rank 1:	.73	.70	.63	.67	.74
Rank 2:	.58	.57	.55	.56	.58
Rank 3:	.49	.48	.48	.48	.49
Rank 4:	.42	.42	.43	.37	.42
Rank 5:	.37	.37	.37	.33	.37
Rank 6:	.32	.32	.33	.29	.32
Rank 7:	.28	.29	.29	.25	.28
Rank 8:	.25	.25	.26	.21	.25

Several answers can be given. First, there has always been concern about this bias. That is why an effort is frequently made to equate (or match) subject groups. Second, because psychologists frequently plan experiments somewhat casually, generally using stylized "cookbook" designs, there has not been the pressure to understand the nature and magnitude of this problem. Randomization and replication have been the methods used to deal with problems created by subject variability although, in fact, randomization does not eliminate the problem at all, and the amount of replication needed to meaningfully reduce the bias threat may often be too costly to employ. In that regard, the experimenter who expends his resources replicating the few-factor experiment to reduce the effect of subject variability is generally exchanging one form of bias for another, i.e., for that which occurs when critical factors are held constant to keep the size of the experiment small.

Third, the holistic approach tends to saturate the experimental design in order to optimize the information-to-cost ratio. This imposes a much greater demand for the cleanliness of each data point. The holistic experiment uses most of its degrees of freedom estimating main- and two-factor interaction effects while tentatively ignoring higher order interaction effects on the working assumption that these latter effects will prove trivial (Simon, 1977b), a matter that will eventually be tested empirically. A few-factor experiment, on the other hand, generally employs a full factorial design that distributes most of its degrees of freedom among the higher-than-two-factor interaction effects and/or replications. For example, with 32 conditions and no replication in either case, a fully saturated holistic experiment could theoretically estimate 16 main effects and 15 strings of two-factor interactions, while a full factorial design could theoretically study five main effects and ten two-factor interactions. The remaining 16 degrees of freedom would be used to estimate higher-than-two-factor interactions. Therefore, under these conditions, in the holistic experiment we can expect some main- and two-factor interaction strings to be biased by the largest expected subject effect while the chances are about half for that to occur with the full factorial. Thus, for any fixed number of estimable effects, the probability of a large subject-bias value being confounded with a critical effect is much greater with a saturated or near saturated design of a holistic experiment than with the full-factorial of the few-factor study.

Finally, holistic experimental designs employ selective confounding to achieve their economy. For example, in the design described in the previous paragraph, all main effects would be aliased (confounded) with three-factor interaction effects. While there exists the calculated risk that this confounding may result in biased estimates of the main effects, in practice the risk is generally low when proper precautions are taken in planning many-factor experiments (Simon, 1977b). Still, when subject effects are also confounded with both higher-order effects and main effects, we only increase the chances that the results will be misinterpreted.

One might ask why the greater concern with subject-factor confounding than with three-factor interaction confounding. The reason is that when confounding occurs between factor-related effects we not only know what effects have been confounded with one another (and should have purposefully selected the combinations), but we also have techniques for isolating them if necessary. When there is bias from subject effects, while we know the approximate magnitude of the bias being distributed among the experimental effects, we do not know with which effects the larger biases will be combined.

Do these weaknesses argue against the use of holistic experiments and for the continued employment of the traditional approach? A detailed response to this is beyond the province of this report. However, it suffices to say that:

Point 5. *Given a specified number of factors to be investigated, and fixed resources with which to investigate them, one will obtain more and better quality information at less cost using the holistic approach than using any series of few-factor experiments* (Simon, 1979).

SECTION III

TECHNIQUES FOR REDUCING RISK OF BIAS FROM SUBJECT/FACTOR CONFOUNDING

We have seen how the severity of the bias resulting from subject-factor confounding is a function of the size of subject variability. Techniques for reducing bias must in one way or another decrease the amount of this variability within the experiment.

While no amount of bias should be acceptable, a "pragmatic empirical" orientation is the cornerstone of the holistic approach. In equipment design research, there are at least two circumstances in which an investigator may not wish to take the additional steps required to reduce this subject-related bias from his results. The first would be when the investigator has a priori knowledge that subject performance variability (on the specific task) is small relative to that of the experimental factors and, therefore, does not justify the expenditure of time and money to reduce it further. The second would be when the investigator is only interested in identifying factor effects that significantly exceed the effects of individual differences.

"ACCEPTABLE" BIAS

How can one determine whether the expected amount of bias is acceptable or not? In statistical terms, this question is asking the experimenter to determine the power of his experiment. Considerations regarding power are too extensive to be treated here and have been covered adequately elsewhere (Cohen, 1969). Basically, we are concerned that large individual differences might, as with any irrelevant source of variance, mask the effects of interest. Just how large an effect must be detected can only be determined by the investigator in the context of the question he seeks to answer in the real world. This in turn will determine how large a subject variance can be tolerated. Quite frequently, a precise measure of subject population variability for the particular task is not available prior to the experiment. Still, if similar studies have been performed, their results may give clues regarding the subject variance. Whenever possible, it is recommended that a rough empirical estimate of between-subject variability be obtained as a standard procedure during the pre-experimental exploratory stage. Also, one may wish to determine whether their subject sample is truly homogeneous, i.e., a single population, by obtaining a measure of both between -- and within -- subject variance on the task in question and seeing if they differ significantly from one another.

Eventually, an estimate of subject variance will be obtained from the sample data after they have been collected. As a characteristic of the sequential strategy of holistic experiments, it is generally wise to conduct a single-replication experiment first and examine the results before collecting more data. In one-subject-per-cell holistic experiments, there are two sources from which the subject variance may be estimated. One such source, found in holistic experiments with a built-in lack-of-fit test (Simon, 1977a, 1977b), would come from the replicated center points, each replication being the performance of a different subject. Another estimate of the "error" variance of the sample (which for all practical purposes can usually be interpreted as subject variance) can be obtained in all 2^{k-p} holistic experiments from the slope of the points in the center portion of the normal order plot. This technique is discussed by Daniel (1959) and Zahn (1975a, 1975b). The use of normal plots are discussed briefly in Appendix C. This information, in conjunction with the rest of the data, may provide the clues needed to decide whether or not additional replications are needed.

In man-machine system research, individual differences in ability -- albeit unmeasured and unmeasurable -- may be a major source of variance, generally a major component of the "error" variance and at times greater than the variance from marginally critical equipment factors (Simon, 1976; Westra, 1982). For this reason, without knowledge to the contrary, the investigator should be prepared to take steps to reduce the bias created by subject-factor confounding, if only as a precautionary measure.

SOME TECHNIQUES FOR REDUCING SUBJECT-RELATED BIAS

There are a number of procedures an investigator may employ to reduce subject variability and, consequently, the risks of subject-related bias. These divide into two classes depending on what is known regarding the subjects' abilities that critically affect task performance. Among these are:

When some information is known regarding subject ability:

1. Treat identifiable and relevant sources of subject variance as factors in the experiment.

2. Use an independent covariate to partial out subject ability.

3. Restrict the subject sample to a highly homogenous set.

4. Use covariate information probabilistically to reduce intepretation errors.

When nothing is known regarding subject ability:

1. Add more subjects per condition.
2. Enlarge the experimental design (new fraction).
3. Avoid a saturated design.

Of course, an appropriate combination of these alternatives will probably be the most successful and economical way to minimize subject variability. The selection of any procedure will depend on a number of considerations, to be discussed as each is examined in more detail.

SECTION IV

HANDLING SUBJECT BIAS WHEN INFORMATION IS AVAILABLE REGARDING ABILITIES

The amount of information one may have regarding the subjects' abilities to perform a given task at a given time can vary considerably. Even with a great deal of effort on the part of the investigator, knowledge of each subject's true ability under the particular set of circumstances demanded for subject-bias reduction is likely to be, at best, marginal. When doing human performance research, it is safer to assume that any independent estimate we may have of subject ability (for a particular task and time period) will be inexact.

For purposes of "handling subject bias" in experiments, only a relative measure of individual ability is needed, i.e., an ordering of subjects according to their ability to perform the task at the time of the experiment. Just how inexact these measures actually are would be represented by the correlation between the scores the subjects have obtained on a selected test of their abilities and the "true" measures of their abilities. Naturally, this correlation will only be an estimate, at best, since had we the true values we would need no other information. Thus, whatever effort we may make to obtain a valid independent estimate of subject ability, we must expect it to be a degraded estimate; this reduces whatever effectiveness that information may have.

The techniques below for reducing subject variability (and the bias from confounding) differ in the amount and/or quality of information required regarding the subjects' abilities. No single technique may be sufficient; the use of one technique does not necessarily exclude the use of another. If they are used judiciously, and in combination, subject bias can probably be reduced to a tolerable level at a reasonable cost. Since, as has been shown, the risk of bias is usually present to some degree in any experiment, these techniques may be applicable to other than holistic experiments whenever different subjects are employed in the different cells.

TECHNIQUE 1: INCLUDE SUSPECTED SOURCES OF SUBJECT VARIABILITY IN THE EXPERIMENT AS FACTORS

Even though we may not be able to quantify individual abilities, we frequently are able to classify subjects into categories that previously have been found to have a critical effect on the performance under investigation. For example, pilots might be divided into those with and those without jet experience, or into groups with different amounts of experience, or different amounts of carrier landing experience, or any combination of these categories along with many other relevant dimensions that are likely to affect pilot performance on a particular task. Subjects performing perceptual motor tasks might be broken into groups on the basis of sex and/or age, since these factors have often been found to result in significant differences in performance.

Thus, if we can identify suspected sources of subject variability that have a better than average chance of affecting the performance under investigation, and make them experimental factors, we will not only have broadened the generalizability of our results, but we will have also reduced the unidentified subject variability and, therefore, the size of the expected bias from subject-factor confounding. Westra (1982, pp. 37-38) used this technique in an experiment to determine how much transfer a number of simulator parameters accounted for when military pilots were being trained to land on an aircraft carrier.

In that experiment, the pilots (subjects) were divided into two groups on the basis of their prior training (a combination of flight hours and experience piloting a particular type of aircraft). This one factor accounted for approximately 20% of the total variance in that experiment. Westra suggested that if even one large effect can be identified that accounts for much of the variability among subjects, "it is reasonable to assume that most of the other estimable effects in the experiment [will be] only trivially biased by subject differences" (p.38). Of course, an investigator is not limited to isolating only one subject factor.

Whether and how to include these identifiable sources of variance in the experiment will not be discussed here. Simon (1977, pp. 53-55) provides some points to be considered when incorporating subject dimensions into holistic experimental designs. Whether or not a subject factor should be a part of the fractional factorial design or introduced orthogonally to the design depends considerably on the factor's complexity and measurability, whether it is quantitative or qualitative, and ultimately the limitations set by time and money.

TECHNIQUE 2: PARTIAL OUT SUBJECT EFFECTS USING A COVARIATE

If it were possible to obtain an independent set of measures representing the relative ability of each subject at the time the critical performance occurred, devoid of differences due to other factors, this could be used to partial out subject ability from the confounded performance scores and leave values that more faithfully represent the effects of the experimental factors. This independent set of measures will be referred to as the "covariate scores" and the test task used to obtain them, the "covariate." A covariate was employed in both the Simon and Roscoe (1981) and Westra (1982) studies to isolate subject effects.* The equation used to partial the covariate score from the corresponding score in the confounded experimental data is:

$$\begin{matrix}\text{Performance score}\\ \text{with covariate}\\ \text{effect removed}\end{matrix} = \begin{matrix}\text{Performance}\\ \text{Score}\\ \text{(X)}\end{matrix} - \left[r_{yx}\left(\frac{sX}{sY}\right) \cdot \begin{matrix}\text{Covariate}\\ \text{Score}\\ \text{(Y)}\end{matrix}\right] \qquad \text{(Eq. 7)}$$

where r_{xy} is the estimated correlation between the performance data and the covariate data. The s stands for the standard deviations of the designated data group, either the performance (X) or the covariate (Y).

Once ability has been partialled out, the remainder scores (which theoretically -- if error variance is trivial -- represent the contributions of the equipment configurations) will be analyzed in the same way the original confounded scores would have been. In any subsequent test of statistical significance, however, one degree of freedom is lost from the error term to cover the use of the covariate.

HOW LARGE SHOULD THE COVARIATE'S VALIDITY COEFFICIENT BE? When covariate scores are partialled from the confounded data, the reduction in bias due to subject variability is directly related to the size of the validity coefficient. Expressed in equation form this is:

$$\begin{matrix}\text{Adjusted bias}\\ \text{(covariate}\\ \text{effectiveness)}\end{matrix} = \text{Original Bias} \cdot \sqrt{1 - r_{yt}^{2}} \qquad \text{(Eq. 8)}$$

where the validity coefficient, r_{yt}, is the correlation between the covariate scores (Y) and the "true" measures of subjects' abilities (T). The reduction in the bias effect as a function of different size validity coefficients are given in Table 6.

*Comments and criticisms of the Simon and Roscoe and the Westra transfer of training experiments can be found in Appendix D along with considerations regarding the development of an adequate covariate for subject bias reduction.

The percent reduction in bias can be calculated for any adjusted bias using this equation:

$$\text{Percent bias reduction} = 1 - \left[\frac{\text{Adjusted Bias}}{\text{Original Bias}}\right] \bullet 100 = \left[1 - \sqrt{1 - r_{yt}^{2}}\right] \bullet 100 \quad \text{(Eq. 9)}$$

TABLE 6. REDUCTION IN EXPECTED BIAS AS A FUNCTION OF THE SIZE OF THE COVARIATE VALIDITY COEFFICIENT (r_{yt})

Validity Coefficient	Percent Reduction in Expected Bias
.15	1.1%
.30	4.6%
.50	13.4%
.65	24.0%
.80	40.0%
.85	47.3%
.90	56.4%

The numbers in Table 6 show that for a covariate to have much effect, its correlation with the "true" measure of subject ability must be quite high, higher than is usually found in practice. If subject bias is to be reduced by one-half, the covariate scores must correlate .866 with the "true" values. If the bias is to be reduced by only one-quarter (i.e., to be 75% of the original), then the correlation must be .661.

Validity coefficients between covariate tests and a criterion task seldom exceed .50 and usually are smaller as the criterion task becomes more complex (e.g., flying an aircraft). Cohen (1969, pp.75-78) provided the following subjective labels -- with qualifications -- for certain correlation values found in the "soft" behavioral sciences: .10 is a "small" correlation; .30 is a "medium" one; and .50 is a "large" one. These numbers are not out of line with the validity coefficients found in many pilot selection tests used today, even those including several covariates.

When one considers the time and cost to develop a covariate that is likely to have at best marginal effectiveness, and realizes that a new one may need to be developed with each moderate shift in the criterion task, one must severely question whether the use of covariates can ever be cost effective.

While psychologists have been content to accept covariates with low validity coefficients when they are used to reduce the size of the error variance, the same levels are not acceptable when the covariates are used to reduce bias. The reason is that even if we are not successful in reducing the "error" variance much by this covariate method, the presence of a large overall error variance at least warns the investigator that the data must be interpreted cautiously. On the other hand, even when we recognize that the data may be heavily biased, we have no way of knowing where the different amounts of bias are distributed. Consequently, we are more likely to draw erroneous conclusions regarding the effects, and in particular, tend to make more Type II errors, eliminating real effects that should have been detected in a screening study.

TECHNIQUE 3: USE MORE HOMOGENEOUS SUBJECTS (BITRUNCATION)

The covariate in Technique 2, to be effective, should be allied as closely as possible to the task and the subjects' ability levels at the time the confounded data were generated (see discussion in Appendix D). Technique 2 is employed at the analysis stage of a research effort after the data have been collected. We may, however, employ covariates in another manner and at another time, i.e., during selection and prior to data collection.

If there are readily available measures likely to be associated with performance on the task, but not suitable for inclusion in the experimental design (Technique 1), then these might be used during subject selection to create a more homogenous group of experimental subjects. Without a costly evaluation of the effectiveness of this material--they might come from personnel records that have a bearing on the task at hand--we may use this material as follows:

1. Select more subjects than you intend to use.

2. Obtain the most relevant data regarding each subject's abilities and order the subjectь accordingly.

3. Truncate the list equally at both ends to arrive at the number of subjects needed for the experiment (e.g., a bitruncation of 30% means that 15% of the subjects were eliminated from each end of the distribution).

Basically, bitruncation of a reasonably normal subject population will result in a reduction of experimental subject variability and a corresponding reduction in bias. The amount of reduction in variability and bias depends on the validity of the data used to rank and select the subjects for truncation. For example, if the correlation between the ranking data and subject ability on the experimental task were zero, no variance and bias reduction would be achieved, even though the subjects with the largest and smallest scores were removed from the sample.

The following equation determines the new expected bias when covariate scores representing subject ability are used to eliminate subjects equally at both ends of the distribution:

$$\underset{\text{(bitruncated)}}{\text{Adjusted bias}} = \text{Original bias} \cdot \sqrt{1 - \left(r_{yt}^{2} \cdot P\right)} \qquad \text{(Eq. 10)}$$

where r_{yt} is the estimated correlation between the covariate scores used to rank the subjects prior to truncation (Y) and the subjects' true abilities (T), i.e., the validity coefficient, and P is the proportion of variance reduction due to bitruncation or P = [1-(new variance due to bitruncation divided by old variance)]. For example, if the variance of the subject population were originally 1, and after bitruncation were .75, then the proportion of variance reduction, P, would be [1 - .75/1.00] = .25. Or if the original variance were 4.50 and the new variance were .90 after truncation, the P would be [1 - .90/4.5] = .80. P is the maximum reduction possible due to bitruncation.

At this time no equation is available for determining the reduction in variance from a normal population as a function of percent bitruncation. Therefore, values for a range of cases were obtained using a Monte Carlo computer simulation. Beginning with a normally distributed population of subject abilities with a variance of one, the distribution of abilities were bitruncated 25%, 50%, and 75%. In these cases, with the original population variance of one, the population variance is reduced to .377, .151, and .041, respectively. Were the square root of these values inserted for the population standard deviation in Equation 4, the reduction in expected bias would be considerable. However, these values are valid only if the covariate used to rank the subjects correlated perfectly with the "true" criterion.

As discussed previously, to assume that the covariate would be this perfect, or anywhere near it, is unreal. Therefore, a correction must be made in the P value for the degradation in the relationship between the ability scores used for truncation and the "true" ones; that is, for the estimated validity coefficient of the covariate. This is taken care of in Equation 10.

The proportion of original bias reduction -- that portion of Equation 10 under the square root sign -- for several percentages of bitruncation and several degrees of validity coefficient degradation are given in Table 7. It must be remembered that we are using bitruncation in lieu of actually knowing how effective our covariate is, without actually having to determine its validity coefficient. It is always better to use the covariate *if* it is known. The term is included in Equation 10 to show relationships and to enable the reader to judge what to expect from bitruncation when he assumes -- from past experience -- how valid his covariate is likely to be.

The data in Table 7 reveals that while bitruncation itself achieves considerable reduction in variance, the validity of the covariate used to rank and eventually select the scores on either end of the distribution severely limits this capability. Since in practice we would be fortunate to find a covariate with a validity coefficient greater than .5, the procedure would appear to have limited utility. Still, since bitruncation without the need to validate a covariate is inexpensive to employ, it can be used. It won't hurt and it may help.

TABLE 7. PERCENT EXPECTED BIAS REDUCTION AS A FUNCTION OF PERCENT BITRUNCATION AND DEGRADED COVARIATE

	Percent Bitruncation*		
	25%	50%	75%
Correlation of covariate with true abilities	Percent Bias Reduction		
.80	22.5	32.4	37.9
.65	14.2	19.9	22.9
.45	6.5	9.0	10.2
.30	2.8	3.9	4.4

*A 25% bitruncation means that 12.5% of the scores have been removed from each end of a normal distribution, leaving the center 75%. The variances of the 25%, 50%, and 75% bitruncated populations, assuming the original N(0,1) population, were .377, .151, and .041, respectively (as estimated from a computer simulation).

It should be noted that a bitruncated sample is, of course, no longer strictly normal nor strictly random and, therefore, would technically place some restrictions on inference procedures. However, in view of what is known about the effect of non-normal populations, and considering the fact that the bitruncation procedure is likely to get rid of "outlier subjects," in practice, it is believed that this concern with non-normality is trivial.

TECHNIQUE 4. USE THE COVARIATE INFORMATION PROBABILISTICALLY TO REDUCE INTERPRETATION ERRORS

The first three techniques all reduce the effects of bias by reducing subject variability. This technique seeks only to reduce the risk of an interpretation error by using covariate information regarding subject ability to identify which factor effects are likely to have been biased by the larger subject bias effects. A variation on this approach was used by Westra (1982) in a multifactor transfer-of-training experiment.

The covariate data regarding subject ability would be used in this way. After subjects have been randomly assigned to experimental conditions, their covariate ability measures would be treated as performance scores and subjected to the usual analysis of variance. This would show for this sample the factor effects with which the largest subject (covariate) bias effects are confounded, but only to the degree that the covariate scores are valid representations of the subjects' true abilities.

When covariate validity is poor, the largest true bias effect may not actually be located where the largest covariate-determined bias effect indicates. Since in a real-world experiment we have only the covariate data to work with, it is valuable to know what the chances are that the larger true bias effects are actually being represented by the larger covariate bias effects.

To obtain this information, the following Monte Carlo computer simulation was performed. Subjects with "true" ability scores drawn from a normal distribution were randomly assigned to the experimental conditions and an analysis of variance was performed on these scores. The factor effects associated with the eight largest positive subject effects were recorded. Then a second set of "covariate" scores based on r_{yt} and normal sampling theory were generated. An analysis of variance was performed on this second set of data (assigned to the same set of conditions as the first), and the eight largest positive covariate subject effects recorded. The matches between the first and second set of data for all combinations of the eight largest effects from each set were noted. This procedure was repeated 10,000 times and summarized for different numbers of effects (K) and for covariates with different validity coefficients (r_{yt}).

The probability that the largest true positive or negative biases would occur where the largest positive covariate biases were located was determined for various combinations, i.e., (1) the probability that the largest true positive or negative bias occurs where the largest corresponding covariate effect is found, (2) the probability that the largest positive or negative true bias will occur among the two to eight largest positive covariate effects, and (3) the probability that multiple sets of the largest positive true biases will occur within multiple sets of positive covariate effects.

In Table 8, for several values of K and r_{yt} , the probabilities that the largest positive true bias is represented by the first or as many as the first eight largest positive covariate bias effects are given. More complete tables are given in Appendix E.

Table 8 is read as follows. For example, when there are 63 effects and the validity coefficient is .30, the probability is only .06 that the largest true positive bias is actually confounded with the factor effect indicated by the location of the largest positive covariate effect. This probability only increases to .32 that the true bias effect will be represented by any one of the first eight positive covariate bias effects. The chances improve considerably when the validity coefficient for the covariate equals .71 (actually .707); in that case, the chances are .5 that the largest true bias effect will actually be confounded with one of the factor effects to which the largest three positive covariate bias effects are confounded.

TABLE 8. PROBABILITY THAT THE LARGEST TRUE POSITIVE SUBJECT BIAS EFFECT WILL BE CONFOUNDED WITH THE FACTOR EFFECTS ASSOCIATED WITH THE LARGEST POSITIVE COVARIATE BIASES

		Number of Largest Positive Covariate Bias Effects Involved							
		1	2	3	4	5	6	7	8
K	r_{yt}	PROBABILITIES							
31	.30	.09	.17	.24	.30	.35	.40	.45	.49
31	.71	.31	.48	.59	.67	.73	.78	.82	.85
63	.30	.06	.11	.16	.20	.23	.26	.29	.32
63	.71	.26	.40	.50	.57	.63	.68	.72	.75
127	.30	.04	.07	.10	.13	.15	.17	.19	.22
127	.71	.22	.34	.42	.49	.54	.59	.62	.66

When one inspects the data in Table 8, as well as those in Table E-1 (Appendix E), certain conclusions can be drawn about the use of covariate subject ability data to identify which factor effects are likely to be severely biased. When all other things are equal, we can say that:

1. The probability that the location of any true bias is likely to be where the covariate biases are located increases as more of the covariate biases are involved.

2. The largest true bias has a greater chance of being found where the largest positive covariate bias is found when fewer effects (K) are being investigated.

3. As the validity coefficient increases, the higher the probability that the true bias and the covariate bias effects will coincide.

None of the above observations is too startling. The most important revelation from Tables 8 and E-1 is the general impracticality of using covariate scores to represent the subject's true ability for purposes of identifying which factor effects are likely to be biased. They can be used, and probably should be used, but only as a precautionary tool. This rather disconcerting condition occurs because in life we seldom find covariate validity coefficients that exceed .50. The usefulness of this technique which depends on covariate data must be considered limited. (See Appendix D.)

But what are the chances that if the covariate measures are used, we might locate a large positive bias where, in fact, the true bias was a large negative one? Error of direction, particularly when larger subject biases are involved, can affect the interpretation of the data. For example, if we see no effect when a large positive one was expected, we might suspect that a large negative bias had cancelled out a large positive effect if we also discover a large negative covariate score could probably be associated with that effect. On the other hand, we might be more willing to accept the lack of an effect when a large positive one had been expected if a large positive covariate score could probably be associated with that location.

Some probabilities regarding directional confusion of bias effects for K=31 and several covariate validity coefficients are shown in Table 9. The probabilities indicate that the chances are relatively small that a large true negative bias would exist where one of the eight largest positive covariate scores were located. For example, with a validity coefficient (r   ) of .30, there is only a .05 chance that the largest true negative bias would be confounded with any factor associated with one of the five largest positive covariate bias effects. As the validity coefficient increases, the chances of such a reversal of direction in the bias effects become even less likely. The same is true as K increases.

TABLE 9. PROBABILITY THAT THE LARGEST TRUE NEGATIVE SUBJECT BIAS EFFECT WILL BE CONFOUNDED WITH THE FACTOR EFFECTS ASSOCIATED WITH THE LARGEST POSITIVE COVARIATE EFFECTS (K=31)

	Number of Largest Positive Covariate Bias Effects Involved								
	1	2	3	4	5	6	7	8	
r_{yt}	Probabilities								
.00	.03	.07	.10	.13	.16	.20	.23	.26	[.06]*
.30	.01	.01	.02	.04	.05	.06	.08	.10	[.01]*
.45	.00	.01	.01	.01	.02	.03	.03	.04	
.65	.00	.00	.00	.00	.00	.00	.01	.01	

*Probabilities in brackets are chances that two true largest negative bias effects will be confounded with the factor effects at which any of the eight largest positive covariate biases are located.

Thus, with only a little information about subject ability, we can reduce our chances of misinterpreting the data with regard to the direction of the largest subject bias effects by a cautious examination of *a priori* expectations about experimental factor effects -- an important step in pre-experimental analysis -- and the probabilities of direction associated with the largest covariate scores.

SECTION V

HANDLING SUBJECT BIAS WITH NO INFORMATION REGARDING ABILITIES

Even when no information is available regarding subject variance, there are still steps we may take to reduce the risks of misinterpretation from subject bias. While we cannot control the variability of our subject sample in this case, we can reduce the bias effects and/or the risk of misinterpretation. These techniques, of course, are also useful when the information regarding subjects' abilities is poor.

The following techniques, while not totally independent, can be used for this purpose:

1. Add more subjects per condition (cell).

2. Avoid saturating the holistic design.

3. Enlarge the experimental design by adding a new fraction.

Let us examine the advantages and disadvantages of each of these.

TECHNIQUE 5: INCREASE THE NUMBER OF SUBJECTS PER CONDITION

The sizes of the expected biases in a specific experiment in which subjects are selected and assigned at random decrease as the number of subjects increases. This relationship is shown by the equation:

$$\text{Adjusted bias} = \text{Original Bias} \cdot \sqrt{\frac{1}{\text{Number of Replicates}}} \qquad \text{(Eq. 11)}$$

where replicates equals the number of subjects per cell.

With 16 experimental conditions and 15 effects, for example, the largest expected bias would be Ø.8679 with one subject per cell (N = 16) but only Ø.6137 with two subjects per cell (N = 32). By doubling the number of subjects, the bias was reduced approximately 29%.

The following equation describes the arithmetic relationship between expected bias reduction and replication:

$$\text{Percent bias reduction (replication)} = \left[1 - \sqrt{1/n}\right] \bullet 100 \qquad \text{(Eq. 12)}$$

where n is the number of times the original design is replicated, i.e., using n different subjects on every experimental condition.

This equation tells us that when the number of subjects per cell are doubled, the size of the biases at all ranks are reduced to .707 of their original sizes. That is a 29.3% reduction. Thus, if we start with one subject per cell, we must increase the number to four subjects per cell before the bias is cut in half. If we had increased from one to three subjects per cell, the biases at all ranks would have been reduced to .577 of their original sizes.

This simple solution to bias -- i.e., increasing the number of subjects per cell -- so flagrantly employed by those doing few-factor experiments -- is not viable in large-scale, many-factored experiments. Since the number of subjects required to fill the cells of even a one-subject-per-cell holistic transfer of training experiment will frequently strain the limit of subject availability in the first place, trying to increase that number may easily exceed that limit. Faced with the prospect of finding more subjects may also cause the experimenter to extend the boundaries of his subject source, thereby increasing subject variance with a less homogeneous sample.

TECHNIQUE 6: KEEP THE HOLISTIC DESIGN UNDERSATURATED

Arithmetically, an unreplicated Resolution IV design of the 2^{k-p} type frequently employed initially in holistic experiments allows a maximum of N/2 main effects to be estimated (independently of any two-factor interaction effects) out of N-1 estimable effects in a design with N conditions. The remaining [N/2 - 1] effects are <u>strings</u> of two-factor interactions (i.e., 2fi) that are aliased with one another within strings, but independent of any other main or two-factor interaction string effects. Main effects are confounded with higher odd-ordered interactions effects and the two-factor interaction strings are confounded with higher even-ordered interactior effects. For example, with 32 conditions, one subject per cell, we are able to estimate 31 effects of which 16 can be main effects clear of two-factor interactions, and the remaining 15 effects being strings of two-factor interactions. With 64 conditions, of the 63 possible effects, we can estimate up to 32 main effects.

When all N/2 main effects are to be studied, the design is said to be "fully saturated."

Insofar as the bias problem is concerned, it is simple arithmetic to see that when given a 32 condition design with one subject per cell, 31 subject effects will be confounded with 31 factor effects. Were we to study only 12 factors instead of the possible 16, for example, with this design our chances are only 12/31 (or 39%) in the unsaturated design versus 16/31 (or 52%) in the saturated design that the largest bias effect will be combined with a main effect. Thus, if bias is an important concern, one trade off requiring no increase in the number of subjects would be to avoid fully saturating the design.

However, the idea of reducing the number of factors being investigated from 16 to 12 to solve a problem is contrary to the holistic philosophy that "all potentially critical factors should be included in the experiment." To make this reduction is to expose oneself to one of more serious criticisms leveled at investigators who do few-factor experiments, who adjust the informational requirements to the design, rather than vice versa. We must find a way of not saturating the design while not sacrificing the inclusion of any potentially critical factors.

A 32-condition design is, at best, marginal in size for any serious holistic experimentation. A 64-condition design is a more comfortable size from the standpoint of flexibility in design and reliability in the analysis. If the design were that large, then we would no longer be deciding between 16 or 12 candidate factors but between 32 or some smaller number. Experience suggests that it is extremely unlikely that there will ever be as many as 32 factors of practical importance in accounting for performance on any specific task. It is more likely that the bulk of the performance variance will be accounted for by a much smaller number of factors. Except in some unusual exploratory effort, it is difficult to imagine that one could not reduce 32 candidate factors to 20 or so after a careful and thorough pre-experiment analysis (Simon, 1977), without doing serious damage to the experiment or to the integrity of the holistic philosophy.

TECHNIQUE 7: EXPAND THE FRACTIONAL FACTORIAL DESIGN

One may enjoy the benefits of both Techniques #5 and #6 and at the same time increase the information in an experiment by doubling the size of the minimum experimental design with a second fraction of the full factorial. This is equivalent to doubling the size of the screening design. Even though we continue to use only one subject per cell, this technique doubles the size of our N. From Equation 12, we know that as long as we continue to study the same number of factors, the expected bias at every rank is decreased proportional to the reciprocal of the square root of the number of times the original design is replicated.

When the number of factors being investigated is fixed, and the same number of subjects is to be employed in either case, an experimenter has to decide whether: he should increase the number of subject per cell n times with a minimum design, or increase the size of the experimental design with d new fractions of the total factorial keeping one subject per cell where n=d. Let us compare the information content and expected bias in the two plans with the following example.

If the same number of factors -- say 12 -- are to be investigated in a typical 32 condition, two subjects per cell, holistic design or a 64 condition, one subject per cell, holistic design, the degrees of freedom in these designs might be partitioned as shown in Table 1Ø.

TABLE 1Ø. COMPARING THE ALLOCATION OF DEGREES OF OF FREEDOM (df) WHEN TECHNIQUES #5 and #6 ARE EMPLOYED

Sources	Design I: 32 cond., 2 S/cell		Design II: 64 cond., 1 S/cell	
	df		df	
Main Effects	12		12	
2-fi strings	15	[5 2fi/per string]	31	[3 2fi/per string]
3-fi strings	4	[14 3fi/per string]	2Ø	[7 3fi/per string]
Within cell	32		Ø	
Total	63		63	

Note: Numbers of 2fi/3fi per string shown here are representative values. Actual values may vary slightly, depending on which columns are involved.

If an investigator decides that every estimable effect (whether in strings or not) should be examined, i.e., 31 for Design I and 63 for Design II, then the largest expected bias effects for a N[Ø,1] population would be .514 and .584, respectively. That is, with the same number of subjects, the biases are somewhat smaller with the smaller design, provided the investigator intends to study all estimable effects. On the other hand, he may seriously only consider the main and two-factor interaction string effects or 27 for Design I and 43 for Design II. In that case, the largest biases would be .5ØØ and .548, respectively, still slightly favoring Design I.

But the size of the bias is not the only criterion for selection, and the differences observed here may be of small practical consequence. There are other considerations.

Thus:

1. TIME AND COST. Since both designs require the testing of 64 subjects, the effort provided to collect the data will be essentially the same for each. The larger design (II) may require some additional time to prepare the hardware and/or software for the additional experimental conditions.

2. INFORMATION. The larger design (II) provides more information about the factors. It enables assumptions regarding some three-factor interactions to be more easily tested.

3. INTERPRETATION CLARITY. The larger design (II) reduces the problems of confounding from two sources: (a) interaction effects, and (b) trends. There are fewer three-factor interactions confounded with one another within each string or with the main effects. Similarly, there are fewer two-factor interactions confounded with one another within each string. This reduction in confounding makes interpretation as well as eventual isolation easier to achieve.

The larger designs, being more robust to linear, quadratic, and cubic trend effects (Simon, 1978), also reduce those sources of confusion. In addition, the larger design allows the investigator greater freedom to assign potentially critical effects to the most trend-resistant columns in the design. By increasing the number of columns in the entire experiment, the investigator has more freedom in assigning factors in a way that facilitates interpretation and reduces the task of later isolating aliased effects (Simon, 1974; 1977).

4. ERROR ESTIMATES. Design I with two subjects per cell provides an independent estimate of the within-cell variance (32 degrees of freedom) which conventionally is used as the "error" term; Design II does not. However, experience has shown that at least half of the estimated effects in Design II (also 32 degrees of freedom) are likely to be chance. Thus, the straight portion of the normal ordered plot will generally provide a tentative approximation of "error variance" (see Appendix C).

5. BIAS. The expected values indicated in Appendix B are valid only when the investigator seriously intends to consider all of the estimable effects. If a priori, through outside knowledge and a careful pre-experiment analysis, he can say he will not be interested in the three-factor interaction strings -- they're error -- or in those two-factor interaction strings that occur in columns that are not trend resistant, then he

reduces the effects being investigated and the sizes of the expected biases. In practice (in any pragmatic empirical approach such as this one), we do not expect even half of the possible effects to be other than error, although we may inspect them all.

In summary, while it is useful to obtain an indication of how large the expected bias might be if it can be realistically determined, and it is imperative that we do whatever is practically possible to reduce it, whether we know what it is or not, we do not want to get caught up in a numbers game in which the presumption of precision is made unrealistically. Ordinarily, the differences between biases .500 and .548 are not likely to be of practical importance in a human performance experiment. On the basis of the information presented above, when feasible, the rule is to opt for the larger design with one subject per cell rather than the smaller design requiring the same N with more subjects per cell.

SECTION VI

CONCLUSIONS

For strategic and economic reasons, data collected in holistic multifactor experiments initially include only one subject per cell. This condition per se, it has been shown in this report, does not pose a serious threat to the effectiveness of the approach. The bias due to subject and configuration confounding is not unique to the holistic approach nor to the testing of a single subject on each configuration. Subject-related bias of this type will be found in any human performance experiment in which different subjects are selected and assigned at random to the different experimental configurations.

It is the "individual differences" (i.e., subject population variance) that produce both bias and variable subject-related error in any experiment. Of these two sources, the bias problem is particularly disturbing in holistic research with its high information-to-cost ratio, for it is more subtle, less identifiable, and more likely to lead to a misinterpretation of the results than would the variable error. While traditional concern has been focused on the variable error, this report was oriented toward the reduction of subject-related bias error.

Since size of the bias is related to the size of the subject variance and the number of subjects in the experiment, efforts to minimize bias and keep the experiment's size within practical bounds must reduce the sample variance in some manner without markedly increasing the number of subjects. Risks of misinterpretation can also be reduced by increasing the ratio between the degrees of freedom in an experimental design and the effects of interest. Frequently the available techniques for minimizing the seriousness of subject-factor confounding must be employed blindly without knowledge of the subjects' true abilities. In this report, a number of such techniques for reducing the bias from subject-configuration confounding were suggested.

While the few-factor experiment can more readily afford to reduce subject bias by employing experimental designs with many replications (i.e., many subject per cell), its limited sampling of the many dimensions of a real-world problem can itself frequently produce other, more insidious biases between the experimental results and equivalent conditions in the real world. Unless the factors being held constant are set at values that place the experimental space within the limits of interest in the real world, the experimental results can be seriously distorted. Generality of experimental results can only occur when the data is sampled over a broad, multifactor space. Once a large number of factors have been identified as critical in a particular investigation, the holistic approach will provide more superior information -- both in quality and quantity -- which will be obtained more economically than is possible with any series of few-factor experiments regardless of the randomized assignment of one subject per cell to help achieve this goal.

REFERENCES

Beyer, W. H. (Ed.) Handbook of tables for probability and statistics. Cleveland, OH: Chemical Rubber Co., 1966.

Birnbaum, A. On the analysis of factorial experiments without replication. Technometrics, 1959, 1, 343-357.

Box, G. E. P., Hunter, W.G., and Hunter, J. S. Statistics for experimenters. New York, NY: John Wiley and Sons, 1978.

Campbell, D. T. and Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago, IL: Rand McNally, 1963.

Caro, P. W. Some factors influencing Air Force simulator training effectiveness. Alexandria, VA: Human Resources Research Organization Technical Report No. 77Ø2, March 1977.

Cohen, J. Statistical power analysis for the behavioral sciences. New York, NY: Academic Press, 1969.

Daniel, C. Use of half-normal plots in interpreting factorial two-level experiments. Technometrics, 1959, 1, 311-314.

Daniel, C. Applications of statistics to industrial experimentation. New York, NY: Wiley, 1976.

Dunlap, W. P. Computer simulation investigating expected size of estimated effects due to sampling error for economic multifactor designs. New Orleans, LA: Tulane University (unpublished paper, 1982).

Harter, H. L. Expected values of normal order statistics. Biometrika, 1961, 48, 151-161.

Hays, W. L. Statistics. New York, NY: Holt, Reinhart, and Winston, 1963.

Orlansky, J. Current knowledge and projection on assessing the effectiveness of training. In Society for Applied Learning Technology, Technology of Training Effectiveness and Evaluation for Productivity Assessment. Warrenton, VA: July 14-15, 1982, pp. 1-12.

Owens, D. B. Handbook of statistical tables. New York, NY: Addison-Wesley, 1962.

Simon, C. W. Economical multifactor designs for human factors engineering experiments. Culver City, CA: Hughes Aircraft Company, Technical Report No. P73-541A, June 1973, 171 pp. (AD A035-108).

Simon, C. W. Methods for handling sequence effects in human factors engineering experiments. Culver City, CA: Hughes Aircraft Company, Technical Report No. P74-541A, December 1974, 197 pp. (AD A035-109).

Simon, C. W. Analysis of human factors engineering experiments: Characteristics, results, and applications. Westlake Village, CA: Canyon Research Group, Inc., Technical Report No. CWS-02-76, August 1976, 104 pp. (AD A038-184).

Simon, C. W. Design, analysis, and interpretation of screening designs for human factors engineering research. Westlake Village, CA: Canyon Research Group, Inc., Technical Report No. CWS-83-77A, September 1977a. (Revised June 1978), 220 pp. (AD A056-985)

Simon, C. W. New research paradigm for applied experimental psychology: A system approach. Westlake Village, CA: Canyon Research Group, Inc., Technical Report No. CWS-04-77A, October 1977b (Revised June 1978), 123 pp. (AD A 056-984).

Simon, C. W. Applications of advanced experimental methodologies to AWAVS training research. Orlando, FL: Naval Training Equipment Center Technical Report No. NAVTRAEQUIPCEN 77=C-0065-1, January 1979, 84 pp. (AD A064-332).

Simon, C. W. and Roscoe, S. N. Application of a multifactor approach to transfer of training research. Orlando, FL: Naval Training Equipment Center Technical Report No. NAVTRAEQUIPCEN 78-C-0060-6, July 1981, 83 pp. (AD A108-499).

Westra, D. P. Simulator design features for carrier landing: II. In-simulator transfer of training. Orlando, FL: Naval Training Equipment Center, Technical Report No. NAVTRAEQUIPCEN 81-C-0105-1, December 1982.

Westra, D. P., Simon, C. W., Collyer, S. C., and Chambers, W. S. Simulator design features for carrier landing: I. Performance experiments. Orlando, FL: Naval Training Equipment Center, Report No. NAVTRAEQUIPCEN 78-C-0060-7, September 1982.

Westra, D. P. Simulator design features for air-to-ground bombing: I. Performance experiment. Orlando, FL: Naval Training Equipment Center, Report No. NAVTRAEQUIPCEN 81-C-0105-4, September 1983.

Williams, A. C., Jr. and Adelson, M. Some considerations in deciding about the complexity of flight simulators. Lackland AFB, TX: Research Bulletin No. AFPTRC-TR-54-106, December 1954.

Zahn, D. A. Modifications of and revised critical values for the half-normal plot. Technometrics, 1975a, 17, 189-200.

Zahn, D. A. An empirical study of half-normal plots. Technometrics, 1975b, 17, 201-211.

APPENDIX A

COMPUTER PROGRAMS

The computer programs presented here were used to identify and verify various relationships described in this report. These programs were written in standard FORTRAN IV to run on a VAX 11/78Ø computer.

Program A-1, written by William Dunlap and Susan G. Brown, Department of Psychology, Tulane University, New Orleans LA, 7Ø118, computes a direct evaluation of equations for expected values of normal order statistics as well as their variances and the expected percent-variance-accounted-for for the ith ordered value of a sample of size N drawn at random from a $\underline{N}[Ø,1]$ population. The equations on which this program was based can be found in Beyer (1966, pp. 258 and 26Ø). The values obtained with this program can be used to calculate expected biases for any size experimental design of the 2^{k-p} class discussed in this report. The values may also be used in the construction of normal plots. The program is accurate to at least five places and was used to generate the data in Tables 3, 4, 7, and 1Ø.

Program A-2 was written by Daniel P. Westra (modifying material submitted by William Dunlap) and provides a Monte Carlo-type simulation to estimate expected values of ordered mean differences and related statistics for 2^{k-p} experimental designs with one or more subjects per cell. This program was used to determine the effects of non-normality on expected biases shown in Table 5 in the body of this report. In that case, a finite discrete population was defined using the order statistic expected values computed by program A-1 with "N" equal to the number which must eventually be drawn to fill an experimental design. These values were sampled at random and replaced prior to the next selection. Since the order statistic values from the normal population with variance of one do not actually have a variance of one, the values were adjusted to make their variance equal one. Results from this program are accurate to at least two decimal points. This program also was used to produce Tables 8, 9, and E-1.

There are several shorter programs and algorithms available for computing normal-order statistics for which some accuracy may be sacrificed. An algorithm with an accuracy sufficient for many purposes is used in the BMDP program P5D. There are also tabled values of order statistics from a normal population for a limited number of "N"s in Beyer (1966), Harter (1961), and Owens (1962).

PROGRAM A-1

EXPECTED NORMAL SCORES, STD. DEVIATION, AND PROPORTION OF VARIANCE.

```
C
C       MAIN PROGRAM
C
        DIMENSION EN(127),EN2(127),PVAR(127)
        TYPE 100
100     FORMAT(' ENTER N ')
        ACCEPT *,N
        SST = 0.
        DO 200 I=1,N
          CALL SCOR(I,N,SN,VAR)
          EN(I) = SN
          EN2(I) = VAR**.5
          SST = SST + (SN*SN)
200     CONTINUE
        TPVAR = 0.
        DO 300 I=1,N
        PVAR(I) = (EN(I)**2)/SST
300     TPVAR = TPVAR+PVAR(I)
        WRITE(3,325)
325     FORMAT(' RANK       E(Xi)          STD. DEV.     PROP. VAR')
        WRITE(3,350)
350     FORMAT('          ____________________________________')
        DO 400 I=1,N
400     WRITE(3,500)I,EN(I),EN2(I),PVAR(I)
500     FORMAT(1X,I3,3F13.5)
        WRITE(3,550)
550     FORMAT('          ____________________________________')
        WRITE(3,600)SST,TPVAR
600     FORMAT(2X,F15.5,11X,F15.5)
        END
C
C
        SUBROUTINE SCOR(J,N,SCORE,VAR)
C
C       COMPUTES THE E(x) OF THE JTH OF N RANKED NORMAL
C       SCORES.
C          SCORE RETURNS WITH THE EXPECTED NORMAL SCORE.
C          VAR RETURNS WITH THE VARIANCE OF THE NORMAL SCORE.
C       COMPUTE THE CONTENT OF THE INTEGRAL LOG FORM
C       C = FACL(N)-FACL(J-1)-FACL(N-J)
C
C       GET THE APPROX. NORMAL SCORE
C
        AN = SCR(J,N)
        ANN = AN
3       ANN = ANN-.1
        IF (FUN(ANN,C,J,N) .GT. 0.) GO TO 3
        RBOT = ANN
4       ANN = ANN+.1
        IF (FUN(ANN,C,J,N) .GT. 0.) GO TO 4
        RTOP = ANN
        RNG = AN-RBOT
        XU = AN
        FU = XU*FUN(XU,C,J,N)
        W = RNG/100.
        PRT1 = 0.
        PRT1V = 0.
C
C       START THE FIRST SIMPSON INTEGRATION
C
10      XL = XU-W
        XM = (XU+XL)/2.
        FM = XM*FUN(XM,C,J,N)
        FL = XL*FUN(XL,C,J,N)
        P1 = PRT1+(FL+4.*FM+FU)*W/6.
        PRT1V = PRT1V+(XL*FL+4.*XM*FM+XU*FU)*W/6.
        IF (PRT1 .EQ. P1) GO TO 15
        W = W*1.1
        XU = XL
        FU = FL
        PRT1 = P1
        GO TO 10
15      CONTINUE
        RNG = RTOP - AN
        XL = AN
        FL = XL*FUN(XL,C,J,N)
        W = RNG/100.
        PRT2 = 0.
        PRT2V = 0.
C
C       START SECOND SIMPSON INTEGRATION
C
20      XU = XL+W
        XM = (XU+XL)/2.
        FM = XM*FUN(XM,C,J,N)
        FU = XU*FUN(XU,C,J,N)
        P2 = PRT2+(FU+4.*FM+FL)*W/6.
        PRT2V = PRT2V+(XU*FU+4.*XM*FM+XL*FL)*W/6.
        IF (PRT2 .EQ.P2) GO TO 25
        W = W*1.1
        XL = XU
        FL = FU
        PRT2 = P2
        GO TO 20
25      CONTINUE
        SCORE = PRT1+PRT2
        SCR2 = PRT1V+PRT2V
        VAR = SCR2-SCORE*SCORE
        RETURN
        END
C
        FUNCTION FUN(X,C,J,N)
C
C       COMPUTES VALUE OF FUNCTION TO BE INTEGRATED, FIRST IN LOG
C       FORM, THEN IN DECIMAL FORM IF LARGE ENOUGH.
C       C = N!/(J-1)!/(N-J)! - THE CONSTANT OF INTEGRATION
C       FUN = C*P**(N-J)*Q**(J-1)*Z
C       WHERE P IS THE PROPORTION BELOW & Q IS THE PROPORTION ABOVE
C       X ON THE NORMAL CURVE; AND Z IS THE ORDINATE OF THE NORMAL
C       CURVE AT X.
C
        DOUBLE PRECISION X,POFZ,P,Q,XJ,XN,FUN,FL
        XJ = J
        XN = N
        P = POFZ(X)
        Q = 1.D0-P
        FUN = 0.D0
        IF (P .LE. 0.D0 .OR. Q .LE. 0.D0)RETURN
        FL = C+DLOG(P)*(XN-XJ)+DLOG(Q)*(XJ-1.D0)-X*X/2.D0-0.918938533D0
        IF (FL .LT. -40.)RETURN
        FUN = DEXP(FL)
        RETURN
        END
```

(Continued)

(Program A-1 continued)

```
C
      FUNCTION FACL(K)
C
C     COMPUTES THE NATURAL LOG OF K!
C     ABOVE 36 USES STIRLINGS APPROXIMATION
C
      IF (K .GT. 36) GO TO 20
      FACL = 0.
      IF (K .LE. 1)RETURN
      DO 10 I=2,K
        X = I
10    FACL = FACL+ALOG(X)
      RETURN
20    X = K
      FACL = .918938533+ALOG(X)*(X+.5)-X+1./(12.*X)
      RETURN
      END
C
C
      FUNCTION SCR(J,N)
C
C     RAPID APPROXIMATION TO THE JTH OF N EXPECTED NORMAL SCORES.
C     USES BLOMS ALGORITHM WITH CORRECTIONS PROPOSED BY
C     HARTLEY, BIOMETRIKA, 1961, 48, 151-165.
C
      XJ = J
      XN = N
      X = ALOG10(XN)
      IF (N .GT. 400) GO TO 10
      A1 = .315065+.057974*X-.009776*X*X
      A2 = .327511+.058212*X-.007989*X*X
      GO TO 20
10    A1 = .3752+.00976*X+.00008*X*X
      A2 = .3866+.01418*X+.00029*X*X
20    A = A2
      IF (J .EQ.1.D0 .OR. J .EQ. N) A=A1
      SCR = ZINV((XJ-A)/(XN-2.*A+1.))
      SCR = -SCR
      RETURN
      END
C
C
      FUNCTION ZINV(P)
C
C     QUICK APPROXIMATION TO THE INVERSE OF THE NORMAL DIST. FUNCTION
C     EMERSON, BRMI, 1979,11, 397-398.
C
      SGN = -1.
      Q = P
      IF (Q-.5)2,2,1
1     SGN = 1.
      Q = 1.-Q
2     Z = SQRT(-2.*ALOG(Q))
      W = Z*(.18926+Z*.001308)
      W = 1.+Z*(1.432788+W)
      Q = .802853+Z*.010328
      W = (2.515517+Z*Q)/W
      ZINV = (Z-W)*SGN
      RETURN
      END
C
C
```

```
      DOUBLE PRECISION FUNCTION POFZ(XX)
C
C     NORMAL PROBABILITY INTEGRAL, -INFINITY TO X; X>0
C     ALGRORITHM 26.2.11 P.932 ZELEN & SEVERO
C
      DOUBLE PRECISION X,XX,S,T,C,XN,SN
      X= DABS(XX)
      POFZ = 1.D0
      IF (X .GT. 8.35) GO TO 20
      S = X
      T = X
      C = X*X
      XN = 0.D0
5     XN = XN+1.D0
      T = T*C/(2.D0*XN+1.D0)
      SN = S+T
      IF (SN .EQ. S) GO TO 10
      S = SN
      GO TO 5
10    POFZ = .5D0+.398942280401432700*DEXP(-C/2.D0)*S
20    IF (XX .LT. 0.D0) POFZ=1.D0-POFZ
      RETURN
      END
```

PROGRAM A-2

```
      EXPECTED MEAN DIFFERENCES OF LARGEST, NEXT LARGEST, ETC.,
      EFFECTS FROM ONE OR MORE SUBJECT PER CELL, TWO LEVEL FAC-
      TORIAL DESIGNS.  SUBJECTS ARE DRAWN FROM BINOMIAL, RANK
      ORDER, AND NORMAL DISTRIBUTIONS WITH TRUE FACTOR EFFECTS
      EQUAL TO ZERO.  THE ONLY SOURCE OF TRUE VARIANCE IS BETWEEN
      SUBJECTS.  PROGRAM CAPACITY IS DESIGNS UP TO 128 CONDITIONS
      AND UP TO 128 SUBJECTS.

      IMPLICIT REAL*4 (M)
      DIMENSION D(3,128),E(3,127),P(128,127),EN(127),ER(127),EB(127),
     X NS(128)
      DIMENSION ESS(3,127),ALOW(3,127),HIGH(3,127),RN(128)
C
C     SIGN MATRIX USED AS BASIS FOR GENERATING OTHER SIZE FACTORIALS.
C
      DATA ((P(I,J),J=1,3),I=1,4) /-1.,-1.,1.,1.,-1.,-1.,-1.,1.,
     X -1.,1.,1.,1./
C
      NR=10000
      SEED=SECNDS(0.0)
C
C     NMAX MUST BE INPUT BY USER . IT SPECIFIES SIZE OF DESIGN
C     ACCORDING TO 2**NMAX.
C     NSCELL ALSO USER INPUT. GIVES NO. SUBS PER CELL.
C
      NMAX=5
      NCMAX=NMAX-1
      NFCON=2**NMAX
      NFEFF=NFCON-1
      NSCELL=1
C
C     DEFINE BINOMIAL AND RANK ORDER SUBJECT DISTRIBUTIONS
C     TOTAL NUMBER OF "SUBJECTS" IS NFCON*NSCELL
C
      NTEMP=NFCON*NSCELL
C
C     RANK DIST(RANGE = 0-1 REGARDLESS OF SIZE)
C
      DO 12 J=1,NTEMP
   12 D(2,J)=FLOAT(J-1)/FLOAT(NTEMP-1)
C
      JJ=NTEMP/2
      JJ1=JJ+1
C
C     BINOMIAL DIST
C
      DO 13 J=1,JJ
   13 D(1,J)=-1.0
      DO 14 J=JJ1,NTEMP
   14 D(1,J)=1.0
C
C     GENERATE THE DESIGN MATRIX
C
      DO 101 N1=2,NCMAX
      NC1=2**N1
      NC2=NC1*2
      NST=NC1+1
      NE1=NC1-1
      NE2=NC2-1
C
C     DUPLICATE EXISTING MATRIX
C
      DO 11 I=NST,NC2
      DO 11 J=1,NE1
   11 P(I,J)=P(I-NC1,J)
C
C     GENERATE NEXT FACTOR
C
      DO 15 I=1,NC1
   15 P(I,NC1)=-1.
      DO 16 I=NST,NC2
   16 P(I,NC1)=1.
C
C     MULT TERMS BY NEW FACTOR
C
      DO 31 I=1,NC2
      DO 31 J=NST,NE2
   31 P(I,J)=P(I,NC1)*P(I,J-NC1)
  101 CONTINUE
C
      DO 20 I=1,NTEMP
   20 NS(I)=I
      DO 21 I=1,3
      DO 21 J=1,NFEFF
      ESS(I,J)=0.
   21 E(I,J)=0.
      DO 22 I=1,3
      DO 22 J=1,NFEFF
      ALOW(I,J)=100.
   22 HIGH(I,J)=-100.
C
C     START OF LOOP FOR THE 10K(OR WHATEVER) RUNS
C
      DO 100 KK=1,NR
C     RANDOM DRAW FROM N(0,1)
C
      CALL RNORM01(NTEMP,RN,SEED)
      DO 509 I=1,NTEMP
  509 D(3,I)=RN(I)
C
  631 DO 25 I=1,NFEFF
      EB(I)=0.
      ER(I)=0.
   25 EN(I)=0.
C
C     RANDOM ASSIGNMENT OF "SUBJECTS" TO CONDITIONS
C
      NTEM2=NTEMP-1
      DO 30 I=1,NTEM2
      IP=NTEMP-I+1
      POS=IP
      RAN = MTH$RANDOM(SEED)
      ICM=RAN*POS+1
      IS=NS(IP)
      NS(IP)=NS(ICM)
   30 NS(ICM)=IS
C     NUMBERS REPRESENTING SUBJECTS MEAN LEVEL INPUT AND TALLIED
C
      DO 40 J=1,NFEFF
      K=1
      DO 40 I=1,NFCON
      A=0.
      B=0.
      C=0.
```

(Continued)

(Program A-2 continued)

```
C     PROVISION TO PUT IN MORE THAN 1 SUB PER CELL.
      K2=K+NSCELL-1
      DO 43 K1=K,K2
      I1=NS(K1)
      A=A+D(1,I1)
      B=B+D(2,I1)
 43   C=C+D(3,I1)
      A=A/NSCELL
      B=B/NSCELL
      C=C/NSCELL
      K=K2+1
      EB(J)=EB(J)+P(1,J)*A
      ER(J)=ER(J)+P(1,J)*B
 40   EN(J)=EN(J)+P(1,J)*C
C
      JJ2=NFCON/2
      DO 41 I=1,NFEFF
C      USE ABS VALUES OF MEAN DIFFERENCES HERE IF DESIRED.
      EB(I)=(EB(I))/FLOAT(JJ2)
      EN(I)=(EN(I))/FLOAT(JJ2)
 41   EN(I)=(EN(I))/FLOAT(JJ2)
C
C     SORT EFFECT SIZES
C
      CALL SORT(EB,NFEFF)
      CALL SORT(ER,NFEFF)
      CALL SORT(EN,NFEFF)
C
C     GET RANGE OF LARGEST DIFF VALUES
C
      DO 42 J=1,NFEFF
      IF(EB(J) .LT. ALOW(1,J)) ALOW(1,J)=EB(J)
      IF(ER(J) .LT. ALOW(2,J))  ALOW(2,J)=ER(J)
      IF(EN(J) .LT. ALOW(3,J)) ALOW(3,J)=EN(J)
      IF(EB(J) .GT. HIGH(1,J)) HIGH(1,J)=EB(J)
      IF(ER(J) .GT. HIGH(2,J)) HIGH(2,J)=ER(J)
 42   IF(EN(J) .GT. HIGH(3,J)) HIGH(3,J)=EN(J)
C
C     SUM UP FOR MEANS AND VARIANCES
C
      DO 45 I=1,NFEFF
      E(1,I)=E(1,I)+EB(I)
      E(2,I)=E(2,I)+ER(I)
      E(3,I)=E(3,I)+EN(I)
      ESS(1,I)=ESS(1,I)+EB(I)**2
      ESS(2,I)=ESS(2,I)+ER(I)**2
 45   ESS(3,I)=ESS(3,I)+EN(I)**2
 100  CONTINUE
C
C     GET MEANS AND VARIANCES
C
      REP=NR
      DO 120 I=1,3
      DO 120 J=1,NFEFF
      E(I,J)=E(I,J)/REP
 120  ESS(I,J)=ESS(I,J)/REP-E(I,J)**2
C
      WRITE(6,753) NSCELL
 753  FORMAT(' NUMBER OF SUBJECTS PER CELL IS',I3)
      WRITE(6,755) NR
 755  FORMAT(' NUMBER OF RUNS IS',I6)
      WRITE(6,780) NFEFF
 780  FORMAT(' NO. EFFECTS IN THIS DESIGN IS ',I5)
      WRITE(6,780)
 700  FORMAT('      EXPECTED MEANS BY DISTRIBUTION',//)
      WRITE(6,756)
 756  FORMAT('        BIN(-1,1)     RANK(0-1)    NORM(0,1)')
      DO 130 J=1,NFEFF
 130  WRITE(6,550) (E(I,J),I=1,3)
 550  FORMAT(1X,3F12.4)
      WRITE(6,701)
 701  FORMAT('1       EXPECTED VARIANCES BY DISTRIBUTION',//)
      DO 140 J=1,NFEFF
 140  WRITE(6,550) (ESS(I,J),I=1,3)
      WRITE(6,702)
 702  FORMAT('1      LOWEST OBTAINED VALUES BY DISTRIBUTION',//)
      DO 150 J=1,NFEFF
 150  WRITE(6,550) (ALOW(I,J),I=1,3)
      WRITE(6,703)
 703  FORMAT('1      HIGHEST OBTAINED VALUE BY DISTRIBUTION',//)
      DO 160 J=1,NFEFF
 160  WRITE(6,550) (HIGH(I,J),I=1,3)
 161  STOP
      END
      SUBROUTINE SORT(X,N)
      DIMENSION X(127)
      M=N
 1    K=0
      DO 2 I=2,M
      J=I-1
      IF(X(J) .LE. X(I)) GO TO 2
      K=J
      S=X(I)
      X(I)=X(J)
      X(J)=S
 2    CONTINUE
      IF(K .EQ. 0)RETURN
      M=K
      GO TO 1
      END
      SUBROUTINE RNORM01(N,X,SEED)
C
C     STUFF TO GENERATE RANDOM NORMAL NUMBERS FROM UN. RANDOM
C     NUMBERS BY BOX AND MULLER(1958).METHOD. IF R1 AND R2 ARE
C     2 UN. RAND NUMBERS, THEN
C            Z1=SQRT(-2.*LOG(R1))*COS(2*PI*R2)
C            Z2=SQRT(-2.*LOG(R1))*SIN(2*PI*R2)
C     ARE A PAIR FORM N(0,1).
C
      IMPLICIT REAL*4 (M)
      DIMENSION X(128)
      DO 50 I=1,N,2
      R1=MTH$RANDOM(SEED)
      R2=MTH$RANDOM(SEED)
      X(I)=SQRT(-2.*LOG(R1))*COS(6.2831853*R2)
 50   X(I+1)=SQRT(-2.*LOG(R1))*SIN(6.2831853*R2)
      RETURN
      END
```

For further information regarding these programs and techniques the reader may contact Daniel P. Westra, Essex Corporation, 1Ø4Ø Woodcock Road, Suite 227, Orlando FL 328Ø3.

APPENDIX B

EXPECTED VALUES, STANDARD DEVIATIONS, AND PROPORTION-OF VARIANCE-ACCOUNTED-FOR

Computer program A-1 supplied in Appendix A was used to compute the expected values for z (the standardized bias score), for s (standard deviation for each bias score), and the proportion-of-total-variance-accounted-for at each of K ranks. These are given for "N" = 31, 63, and 127 in Table B-1 that follows. The values are accurate to five places. For "N" = 15 see Table 3, p. 19.

The notation "N" is used to more directly relate these tables to similar ones found in other sources where the notation N is used for a number of cases, rather than C or K. In the context of experimental results, as used here however, the odd-value tables with "N" equal to 31, 63, and 127 would ordinarily be used for determining the expected mean of biases of the K effects in experimental designs in which C and N are 32, 64, and 128. Still, there is no mathematical reason why the tables supplied here need be limited to "experimental effects."

Proportion-of-variance-accounted-for at any rank can be obtained by squaring the expected z value at that rank and dividing by the total sum of squares (TSS) of the z values, following Equation 6 on page 21 . The TSS for each set of expected values is given at the bottom of the appropriate subtable. A period (.) is placed below the rank at the point where the cumulative proportion of variance exceeds 0.30 for the positive z-scores.

Other tables for "N" values less than 89 can be found in Harter (1961). This is the only known table for "N" = 127.

TABLE B-1. EXPECTED VALUES FOR ORDER STATISTICS

"N" = 31

RANK	Z-VALUE	STD. DEV.	PROP. VAR
1	2.05647	0.49362	0.14813
2	1.63167	0.36880	0.09326
3	1.38269	0.31944	0.06697
4	1.19804	0.29170	0.05028
5	1.04709	0.27363	0.03840
6	0.91689	0.26083	0.02945
7	0.80065	0.25133	0.02245
8	0.69438	0.24406	0.01689
9	0.59545	0.23841	0.01242
10	0.50206	0.23398	0.00883
11	0.41287	0.23053	0.00597
12	0.32596	0.22789	0.00374
13	0.24322	0.22593	0.00207
14	0.16126	0.22458	0.00091
15	0.08037	0.22379	0.00023
16	0.00000	0.22353	0.00000
17	-0.08037	0.22379	0.00023
18	-0.16126	0.22458	0.00091
19	-0.24322	0.22593	0.00207
20	-0.32686	0.22789	0.00374
21	-0.41287	0.23053	0.00597
22	-0.50206	0.23398	0.00883
23	-0.59546	0.23841	0.01242
24	-0.69438	0.24406	0.01689
25	-0.80066	0.25133	0.02245
26	-0.91688	0.26083	0.02945
27	-1.04709	0.27363	0.03840
28	-1.19804	0.29171	0.05028
29	-1.38269	0.31944	0.06697
30	-1.63167	0.36880	0.09326
31	-2.05647	0.49362	0.14813

TSS = 28.54883

"N" = 63

RANK	Z-VALUE	STD. DEV.	PROP. VAR
1	2.33781	0.45186	0.09054
2	1.95626	0.32967	0.06340
3	1.73905	0.28116	0.05010
4	1.58178	0.25351	0.04145
5	1.45605	0.23503	0.03512
6	1.34981	0.22180	0.03018
7	1.25698	0.21161	0.02617
8	1.17388	0.20356	0.02283
9	1.09820	0.19694	0.01998
10	1.02833	0.19148	0.01752
11	0.96317	0.18683	0.01537
12	0.90188	0.18283	0.01347
13	0.84380	0.17939	0.01179
14	0.78844	0.17639	0.01030
15	0.73539	0.17378	0.00896
16	0.68436	0.17145	0.00776
17	0.63504	0.16941	0.00668
18	0.58724	0.16757	0.00571
19	0.54073	0.16598	0.00484
20	0.49536	0.16456	0.00406
21	0.45101	0.16330	0.00337
22	0.40753	0.16218	0.00275
23	0.36480	0.16121	0.00220
24	0.32273	0.16036	0.00173
25	0.28122	0.15963	0.00131
26	0.24019	0.15902	0.00096
27	0.19957	0.15850	0.00066
28	0.15927	0.15809	0.00042
29	0.11923	0.15777	0.00024
30	0.07938	0.15754	0.00010
31	0.03966	0.15741	0.00003
32	0.00000	0.15736	0.00000
33	-0.03966	0.15741	0.00003
34	-0.07938	0.15754	0.00010
35	-0.11923	0.15777	0.00024
36	-0.15927	0.15809	0.00042
37	-0.19957	0.15850	0.00066
38	-0.24019	0.15902	0.00096
39	-0.28122	0.15963	0.00131
40	-0.32273	0.16036	0.00173
41	-0.36480	0.16121	0.00220
42	-0.40753	0.16218	0.00275
43	-0.45101	0.16330	0.00337
44	-0.49537	0.16456	0.00407
45	-0.54073	0.16598	0.00484
46	-0.58724	0.16758	0.00571
47	-0.63504	0.16941	0.00668
48	-0.68436	0.17144	0.00776
49	-0.73539	0.17370	0.00896
50	-0.78844	0.17638	0.01030
51	-0.84380	0.17940	0.01179
52	-0.90188	0.18284	0.01347
53	-0.96317	0.18683	0.01537
54	-1.02834	0.19147	0.01752
55	-1.09820	0.19695	0.01998
56	-1.17388	0.20356	0.02283
57	-1.25699	0.21160	0.02617
58	-1.34981	0.22180	0.03018
59	-1.45605	0.23502	0.03512
60	-1.58179	0.25350	0.04145
61	-1.73905	0.28116	0.05010
62	-1.95626	0.32968	0.06340
63	-2.33781	0.45186	0.09054

TSS = 60.36558

Table B-1 (Continued)

"N" = 127

RANK	Z-VALUE	STD. DEV.	PROP. VAR
1	2.59185	0.41902	0.05408
2	2.24250	0.29997	0.04049
3	2.04752	0.25275	0.03375
4	1.90851	0.22545	0.02932
5	1.79873	0.20737	0.02605
6	1.70598	0.19476	0.02346
7	1.62778	0.18469	0.02133
8	1.55775	0.17619	0.01954
9	1.49449	0.16979	0.01798
10	1.43670	0.16441	0.01662
11	1.38340	0.15958	0.01541
12	1.33375	0.15548	0.01432
13	1.28722	0.15174	0.01334
14	1.24333	0.14842	0.01245
15	1.20171	0.14543	0.01163
16	1.16205	0.14288	0.01087
17	1.12413	0.14062	0.01017
18	1.08776	0.13854	0.00953
19	1.05286	0.13638	0.00892
20	1.01915	0.13454	0.00836
21	0.98651	0.13301	0.00784
22	0.95497	0.13133	0.00734
23	0.92427	0.13008	0.00688
24	0.89451	0.12862	0.00644
25	0.86544	0.12751	0.00603
26	0.83714	0.12628	0.00564
27	0.80948	0.12521	0.00528
28	0.78239	0.12430	0.00493
29	0.75589	0.12337	0.00460
30	0.72997	0.12233	0.00429
31	0.70447	0.12155	0.00400
32	0.67941	0.12081	0.00372
33	0.65479	0.12009	0.00345
34	0.63058	0.11936	0.00320
35	0.60670	0.11873	0.00296
36	0.58317	0.11814	0.00274
37	0.55997	0.11756	0.00252
38	0.53704	0.11707	0.00232
39	0.51443	0.11653	0.00213
40	0.49208	0.11601	0.00195
41	0.46994	0.11558	0.00178
42	0.44805	0.11515	0.00162
43	0.42636	0.11477	0.00146
44	0.40489	0.11437	0.00132
45	0.38360	0.11401	0.00118
46	0.36245	0.11371	0.00106
47	0.34150	0.11340	0.00094
48	0.32068	0.11311	0.00083
49	0.30000	0.11285	0.00072
50	0.27947	0.11260	0.00063
51	0.25903	0.11238	0.00054
52	0.23870	0.11218	0.00046
53	0.21847	0.11199	0.00038
54	0.19833	0.11182	0.00032
55	0.17828	0.11167	0.00026
56	0.15829	0.11153	0.00020
57	0.13837	0.11141	0.00015
58	0.11850	0.11131	0.00011
59	0.09868	0.11122	0.00008
60	0.07889	0.11115	0.00005
61	0.05914	0.11110	0.00003
62	0.03942	0.11106	0.00001
63	0:01971	0.11103	0.00000
64	0.00000	0.11103	0.00000
65	-0.01971	0.11104	0.00000
66	-0.03942	0.11106	0.00001
67	-0.05914	0.11110	0.00003
68	-0.07889	0.11115	0.00005
69	-0.09868	0.11122	0.00008
70	-0.11850	0.11131	0.00011
71	-0.13837	0.11141	0.00015
72	-0.15829	0.11153	0.00020
73	-0.17828	0.11167	0.00026
74	-0.19833	0.11182	0.00032
75	-0.21847	0.11199	0.00038
76	-0.23869	0.11218	0.00046
77	-0.25902	0.11238	0.00054
78	-0.27947	0.11260	0.00063
79	-0.30000	0.11286	0.00072
80	-0.32068	0.11312	0.00083
81	-0.34149	0.11341	0.00094
82	-0.36245	0.11371	0.00106
83	-0.38359	0.11402	0.00118
84	-0.40488	0.11438	0.00132
85	-0.42635	0.11478	0.00146
86	-0.44805	0.11515	0.00162
87	-0.46995	0.11557	0.00178
88	-0.49207	0.11602	0.00195
89	-0.51443	0.11652	0.00213
90	-0.53704	0.11706	0.00232
91	-0.55997	0.11757	0.00252
92	-0.58317	0.11815	0.00274
93	-0.60669	0.11875	0.00296
94	-0.63058	0.11935	0.00320
95	-0.65478	0.12010	0.00345
96	-0.67941	0.12081	0.00372
97	-0.70447	0.12155	0.00400
98	-0.72996	0.12236	0.00429
99	-0.75589	0.12337	0.00460
100	-0.78239	0.12431	0.00493
101	-0.80947	0.12523	0.00528
102	-0.83714	0.12628	0.00564
103	-0.86544	0.12751	0.00603
104	-0.89449	0.12867	0.00644
105	-0.92428	0.13005	0.00688
106	-0.95496	0.13137	0.00734
107	-0.98650	0.13306	0.00783
108	-1.01913	0.13459	0.00836
109	-1.05286	0.13638	0.00892
110	-1.08776	0.13858	0.00953
111	-1.12415	0.14055	0.01017
112	-1.16204	0.14294	0.01087
113	-1.20171	0.14543	0.01163
114	-1.24331	0.14847	0.01245
115	-1.28721	0.15176	0.01334
116	-1.33376	0.15542	0.01432
117	-1.38338	0.15965	0.01541
118	-1.43669	0.16446	0.01662
119	-1.49450	0.16972	0.01798
120	-1.55772	0.17630	0.01954
121	-1.62779	0.18467	0.02133
122	-1.70701	0.19465	0.02346
123	-1.79870	0.20747	0.02605
124	-1.90849	0.22554	0.02932
125	-2.04751	0.25276	0.03375
126	-2.24250	0.29997	0.04049
127	-2.59185	0.41902	0.05408

TSS = 124.21255

APPENDIX C

NORMAL ORDER PLOTS

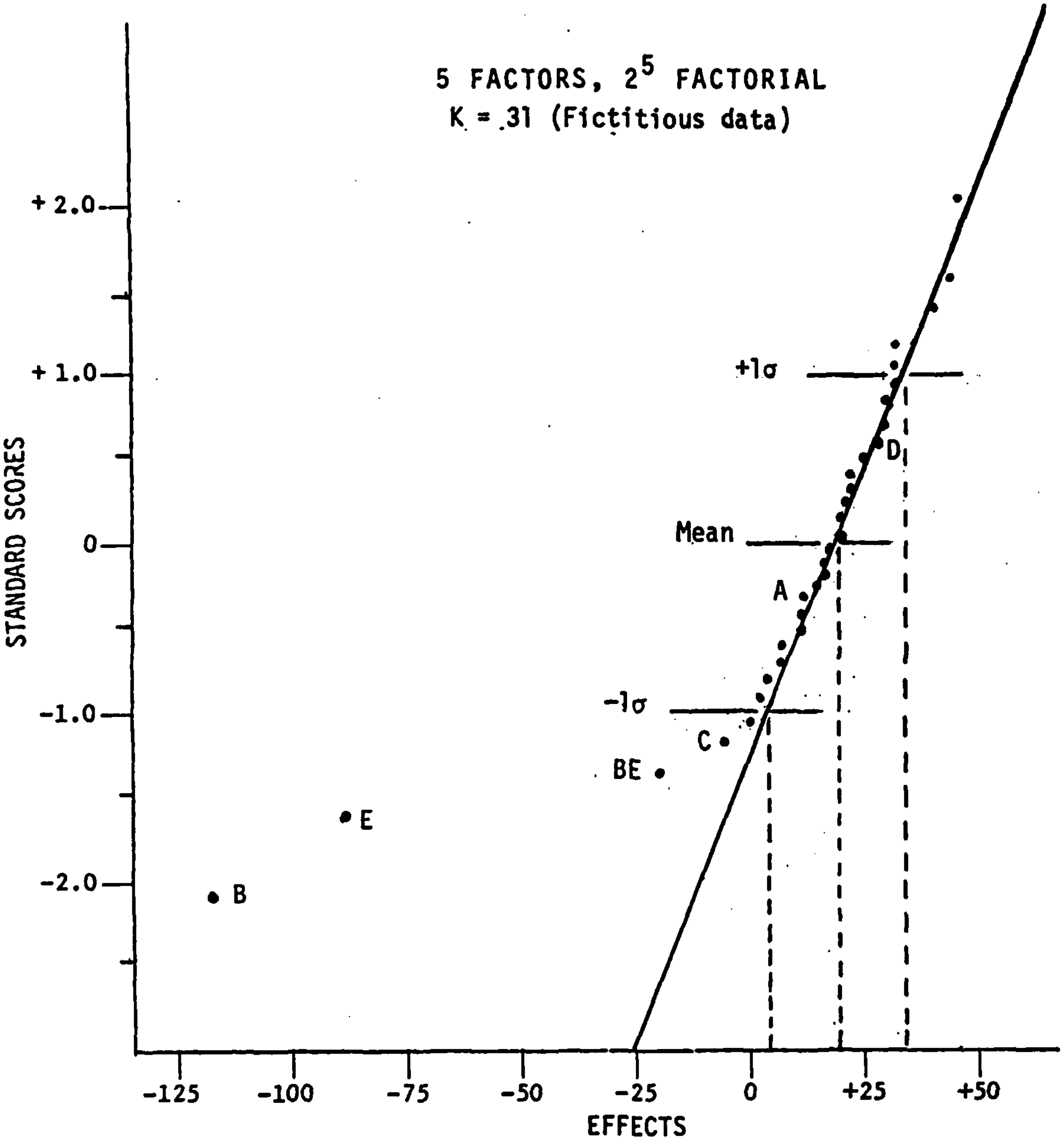
While an extensive discussion of normal-order plotting is beyond the scope of this report, this section will provide those readers unfamiliar with the process with some knowledge of its advantages and disadvantages and how it is used to interpret data from many-factored holistic experiments with one or more subjects per cell.

The mechanics for setting up a plot are as follows: (1) Prepare an effects (i.e., mean differences) scale on the abcissa; (2) mark off the critical probability values on the ordinate using the probabilities equivalent to the z-score values in a table of normal-order statistics for the K effects in the particular analysis; (3) plot the effects (mean differences) from the analysis of variance in order of magnitude, the largest positive effect being at rank one and the largest negative effects being rank K. An example is shown in Figure C-1.

If the observed differences among the effects (mean differences) in the sample are due only to chance, they will fall approximately along a straight line. Even the largest positive and the largest negative effects, usually more than two standard deviations from the mean of all effects, will fall on the line at opposite ends of the plot since deviations that large can be expected by chance.

If one or more effects at either end of the normal plot fall far enough off the line in the appropriate direction, then this suggests that the effects are larger than would be expected by chance. In a screening experiment, when this is observed, the sources of these effects would be subjected to the greatest scrutiny and investigated further.

Confidence limits (called "guardrails") at each rank can be assigned, based on the expected normal distribution around the expected values for the normal-order plots and the number of effects being examined. The probability values for these guardrails can be adjusted to reflect the level of confidence the investigator wishes to use to avoid Type I and Type II errors. At this time, two papers by Zahn (1975a, 1975b) have the best discussion regarding the construction of guardrails for half-normal plots; this can be generalized to the case when normal plotting is employed.



Interpretation:
Factors B and E and Interaction BE are critical. Factor C is marginal; more data is needed to be certain. Factors A and D had no effect.

Figure C-1. Example of an ordered normal plot.

NORMAL OR HALF-NORMAL PLOTS?

In his early work, Daniel (1959) proposed that "half-normal" plots be used rather than normal plots for these purposes. With half-normal plots, the signs of the mean differences are ignored when the data are plotted; with normal plots, the signs are retained. Both forms of plotting are based on the same principles. The normal plot used the total probability scale from minus infinity to plus infinity, while the half-normal plot uses only half the scale from .5Ø to plus infinity and requires that the probability values at which the data would be plotted be adjusted. In a more recent book, Daniel (1976) expressed the opinion that he preferred to use normal plots, believing that the information provided by the signs made it a more useful technique. Simon (1977a) has summarized Daniel's and Zahn's (1975) papers on the use of half-normal plots.

But whether to use a normal or a half-normal plot to test the statistical significance of an effect is not completely arbitrary. When Daniel used half-normal plots, his examples were mainly ones in which the plot served as a quasi-test of significance for the effects from an analysis of variance. When he used the normal plot, he was actually evaluating residuals. However, a more fundamental basis for selecting one or the other exists.

POINT 7: WHETHER ONE USES NORMAL OR HALF-NORMAL PLOTS DEPENDS ON THE NATURE OF THE HYPOTHESIS BEING TESTED. IF THE HYPOTHESIS IS A TWO-TAILED TEST -- I.E., DOES A DIFFER FROM B REGARDLESS OF DIRECTION -- THEN THE HALF-NORMAL PLOT SHOULD BE USED. IF THE HYPOTHESIS IS A ONE-TAILED TEST -- I.E., IS A LARGER THAN B -- THEN THE NORMAL PLOT IS APPROPRIATE.

While psychologists generally have made more use of two-tailed tests, not necessarily chosen on a rational basis, in most equipment/system design studies we are usually concerned with whether one system or device is better than another, a directional question, in which case a normal plot is more appropriate. On the other hand, if one has no expectations regarding direction, i.e., which level of a factor should be better, then the half-normal plot should be used. In mixed cases, use the normal plot.

When there is concern for subject-related bias, as in this report, it is important to retain the signs of the effects in our analysis. The bias from subject effects will be both positive and negative. It is recommended as part of the pre-experimental analysis to try to arrange the factor levels in the experimental design so that the coded values, +1 and -1, are consistently assigned to the level more likely to yield the larger and smaller performance values, respectively. When results turn out to be the opposite of that which had been anticipated prior to data collection, this warns that a large negative subject bias may have been confounded with that effect. While not definitive, this approach helps facilitate the interpretation of the data.

ADVANTAGES AND DISADVANTAGES

Evaluating experimental results with normal or half-normal plots has advantages and disadvantages. Among the advantages are: (1) A test of statistical significance can be made when no independent source for estimating the error variance is available. (2) An automatic adjustment is made for multiplicity (i.e., examining a great many effects at once). (3) The slope of the line in the middle range of the normal plot or the lower end of the half-normal plot provides a means of estimating the "error" standard deviation. (4) Absurdities in the shape of the plot can warn of peculiarities in the basic data. Normal-order plots can and should be used even when an independent error term is available.

Among the disadvantages of normal-order plots are: (1) The mathematics of normal-order plots for identifying critical factors is not fully developed (Birnbaum, 1959; Zahn, 1975). (2) The experimenter must still make certain subjective judgments and assumptions regarding how many factors he expects might be significant and what a suitable overall significance level should be. However, such problems as these are neither overwhelming nor unique to normal-order plots. Since these multifactor experiments are conducted in the screening stage, one does not expect that all decisions be made with statistical precision. An investigator at this time should lean toward liberal interpretations of the results so as not to discard a potentially critical factor. Cautious interpretation of eyeball examination of the normal plot without guardrails will usually suffice for that purpose.

Though there is still much more to be learned about the effective application and interpretation of normal-order plots, they are already a useful tool for interpreting the results from holistic experiments. Even with all the uncertainties, they probably are more effective than the traditional method used by behavioral scientists to test for statistical significance. As Daniel (1976) wrote: "The standard form of the 'analysis of variance' which is widely used in summarizing factorial designs with factors at many levels does not seem to me to be useful for [2] data. All the contrasts from a [2^k] data set must be examined together. Their order, their distribution, and their signs are all lost in the standard analysis of variance table. The habit of summarizing the results in such a table (manifested in so many textbooks that it would be unkind to name) has had a tranquilizing effect with much information lost" (p. 128).

APPENDIX D

CONSIDERATIONS REGARDING THE USE OF COVARIATES FOR SUBJECT-BIAS REDUCTION

Simon and Roscoe (1981) and Westra (1982) used covariates to partial out of their experimental data the bias that results when subjects of different abilities are randomly assigned to the cells of the experimental design. In retrospect, the adequacy of a covariate approach for this purpose must be questioned. In the main body of this report, it was shown that for a covariate (or battery of covariates) to be more than trivially effective, its validity coefficient, i.e., the correlation of the covariate with the true criterion, must be much higher than has been typically achieved when complex tasks are involved.

Here we will briefly describe what some of the characteristics of covariates for bias reduction must be, distinguish between covariates for that and other purposes, and finally discuss why obtaining an adequate covariate for bias reduction is unlikely to be cost effective.

CHARACTERISTICS OF COVARIATES FOR BIAS REDUCTION

At the end of a transfer-of-training experiment in which a different subject is tested on each training condition, the investigator is faced with a set of performance scores, each representing the combined effect of subject ability at the time the task was performed and the difficulty of the experimental condition. Since the purpose of the experiment is to learn about experimental factor effects, the investigator must find a way to remove the effects due to subject differences from the confounded performance scores. To do this with covariates, he must find an independent set of measures that will reflect the rank order of the subjects' abilities on the task at hand at the time the performance data were collected.

Thus, the two primary characteristics of a viable covariate for bias reduction are:

1. The independent measure of subject ability must be made on a task comparable in complexity and difficulty to the one being investigated and involving essentially the same critical skills.

2. Subjects' abilities must be at essentially the same level they were at the time the critical experimental trials were being performed, not what they might have been before the skills were developed nor necessarily after they were fully developed.

Psychologists have frequently been content to use a task without those characteristics, arguing that the covariate task they are using has been found to correlate significantly with the "true" task (which presumably has those characteristics). Unfortunately, they have often been remiss in two ways, i.e., (1) failing to recognize that correlations that are statistically significant may not be of practical significance, and (2) failing to validate their covariate against a task with those properties.

COVARIATES FOR BIAS REDUCTION VS ERROR REDUCTION

A covariate test suitable for bias reduction will not necessarily be the same as one intended to remove subject differences from the error term so as to improve the sensitivity of the test of statistical significance. While isolating individual differences from the error estimate and reducing the subject-induced bias are both accomplished by the same covariate effort, the criterion for effectiveness will differ.

When purifying the error term, we may be content with the attitude that "any improvement will do;" that attitude is not appropriate where bias is concerned. If a covariate is not terribly effective in reducing the error variance, the visible presence of a large error term should at least make the experimenter more cautious about his interpretation of the data. On the other hand, biases are less visible and only a few of the k effects will be critically confounded with the largest biases. While we may be aware that overall the biases are there, we do not know with which effect each hidden bias is confounded, its direction, nor if it is large enough to matter. Chance combinations of the larger biases with experimental effects in the single experiment can make a marginal effect appear significant or a large effect appear trivial. For these reasons, it is important that the covariates used primarily for bias reduction be more effective than those used primarily for error variance reduction.

COVARIATES FOR BIAS REDUCTION VS SELECTION

A covariate for selection purposes also has different requirements from that for bias reduction. For selection purposes, covariate tests must be administered before the critical period, before skills have been developed to any degree, since its purpose is to predict future performance. This is not the case with the covariate for bias reduction which (theoretically, at least) may be administered at any time provided the essential requirements are met.

If the covariate must be given before a skill has been developed, the selection battery will usually have to be composed of tasks markedly different from (and simpler than) the training and criterion tasks. For effective bias reduction, the more similar the critical elements of the task and the covariate are, the better.

Selection tests ordinarily are intended to predict ultimate ability. For bias reduction, we are concerned with subject ability at some crucial point in the learning curve, not necessarily the final stage.

Finally, regarding certain practical and financial considerations, since selection tests are intended to be used over an extended period of time and with numerous applications, the great deal of time and money required to develop them may be justified. On the other hand, one must weigh the costs when one considers that the covariate for bias reduction may be used only once, normally being experiment-specific.

FINDING A SUITABLE COVARIATE

A covariate test (or battery of tests) may be derived from essentially the same task as that employed in the experiment, or it may be derived from simpler tasks which in aggregate presume to measure skills required to perform the criterion task.

COVARIATE TASK = CRITERION TASK. In the first case, if the covariate task is some variation of the criterion transfer task, then the skills involved and measured in both cases will be essentially the same. Ordinarily, no formal validation effort would be required. As the correspondence between the covariate task and the transfer task begins to diverge, e.g., if only a component or simplified version of the criterion task were used in place of the total task, then some estimate of the validity coefficient is probably necessary.

COVARIATE TASK = SIMULATOR TASK. If the covariate task is measured in the simulator rather than under the operational conditions of the criterion task, a frequent practice, then some empirical evaluation of how faithful the critical elements of the simulator are to the real world seems imperative. (Investigators sometime perform quasi-transfer experiments in which the criterion task is performed on another condition in the training simulator. Under the assumption that that latter task is sufficiently "similar" to the real world task, results are interpreted as if they were equivalent to carrying out the transfer phase of the study under operational conditions. Because the simulator cannot really be used with any confidence as a substitute for the real world without a considerable validation effort, we will not consider that condition in this report.

This requirement again raises in a different context the unresolved question of how much and what kind of fidelity must be built into the simulator. When used as the covariate task, it must be faithful enough to measure most of the critical abilities employed in the criterion task and not require others that are not important operationally.

Unfortunately, such a rule is too vague to provide the information needed by either the engineer who wishes to build the least expensive, most effective simulator, or the experimenter who must decide whether the simulator is an adequate medium for the covariate task. Too often, the faithfulness of the simulator is determined by "face validity." This is not enough; fidelity for both purposes must be empirically demonstrated. To date, the empirical procedure employed for that purpose has been inadequate since it has been limited primarily to measuring overall transfer from simulator to the criterion task. In most cases, although the results have been positive, this procedure will not detect the fact that some critical components of the simulator may actually result in a negative transfer effect which has been hidden by a larger positive transfer effect from the other components. Information regarding component transfer is what our research and simulator evaluation efforts should provide in order to build better simulators and to be able to use them for the covariate task.

When using the simulator task for the covariate, one must use a configuration that was not used during training. The condition at the center of the experimental space was used by the experimenters cited earlier. However, with categorical variables that is not possible. Furthermore, if subjects have been tested at the center of the experimental space to obtain data to evaluate the lack of fit of a linear model, using that same configuration for the covariate test raises questions of the purity of such measures.

COVARIATE TASK = TEST BATTERY. When the covariate task is not some variation of that actually used in the experiment, we distance ourselves still further from the criterion task. While one school of thought among psychologists has been that a battery of simple tasks each measuring a different ability factor can represent more complex tasks, in practice the success of such a philosophy has been marginal at best. The reason for this can be more readily appreciated if one perceives performance on a complex task as the linear sum of the effects of a great many factors and possibly some of their interactions.

$$Y = X1 + X2 + \ldots \ldots Xn + X1X2 + X1X3 \ldots \ldots XmXn$$

each with a coefficient reflecting their relative weights.

While simple tasks may be found requiring the abilities used in some elements of the complex task, still, because of that complexity, it is not likely that enough simple tasks would be found to cover all of the important skills, and particularly their interactions or those required to handle the complexity itself. Certainly the skills which psychologists are able to label are only a part of those required to do the task since they are too frequently selected because they are easy to name and to identify and often lack precise definitions; the more difficult to measure and microscopic features are frequently ignored. Other complexities of interskill relationships, being rationally unfathomable, are seldom considered. As a result, there are many components that will not be measured. The proportion of the total number of components (properly weighted) that are accounted for reflects the size of the validity correlation. The proportion that is not accounted for is associated with the "error" (residual) variance that frequently is much larger than the predicted. In this model, we can see why the more similar the covariate is in content and complexity to the criterion task, the better the predictor it will be. One should not underemphasize the tremendous part of this unwritten and unknown equation that never surfaces. When the factors even in simple experiments fail to account for 40% to 50% of the total variance (Simon, 1976), then it should be evident that what the psychologist is capable of labeling and measuring is usually only the tip of the iceberg.

If the covariate task is a battery of simpler tasks presumed to tap the skills affecting the performance of the criterion task, then empirical verification is imperative. Verification in this case must show that most (all?) of the more critical skills required to perform the operational task are represented and at the skill level present at the time when bias is to be removed. How large a validity coefficient must be to have an effective covariate test or test battery is shown in Table 6 in the main text. Since, to be truly effective, these values must be higher than are usually obtained in practice, one must question the viability of this approach when using a covariate for bias reduction. To add to the difficulty when using indirect covariate tasks is the cost involved: (1) to find enough representative tasks, and (2) to validate them when changes in the experiment change the skill requirements.

While one may argue that _if_ a validation test finds that the two tasks do in fact show a high correlation then one could safely use the simpler task. That may well be, _if_ a high enough relationship can be found and _if_ that correlation is obtained when the subjects' skill levels approximated those expected at the critical period in the experiment. But if history tells us anything, the probability that both of these conditions will be met are low indeed and with the costs of developing and validating covariates, which must be reexamined each time the subject and task characteristics change, the viability of the use of covariates to reduce a bias, except as an adjunct technique, appears moot.

CONSIDERATIONS REGARDING THE TIME WHEN COVARIATE MEASURES ARE TAKEN

When the criterion task is used as the covariate, it must be used after the subjects have been trained sufficiently to meet the skill level requirement cited earlier. This means that these measures should be taken shortly before or after the transfer period. This may contaminate the experimental data in several ways: (1) If taken under operational conditions before the criterion measure is made, but after the training, the effects of training can carry over (transfer) to the covariate measures distorting them. The presence of the covariate between training and transfer could also have unknown carryover effects to the measures of the experimental transfer trials. (2) If taken after the transfer data have been obtained, the covariate measures may be contaminated by transfer effects from the criterion task to the covariate task.

If the covariate test is given in the simulator just after the training sessions, then there could readily be carryover (transfer effects) from training to the covariate task in the same way it is expected to occur to the criterion task. While there is often no good way of knowing whether these dangers occur, it is not unreasonable to believe that they can and do and, therefore, steps should be taken *a priori* to have them reduced or avoided.

If covariate measures are taken in the simulator after the transfer phase is complete, and if the different subjects have shown different transfer effects due to difference in training, then unless one is careful (and can demonstrate that there is no danger), these differences may continue to carry over to the covariate measures. One solution to this problem might be to run a number of buffer trials between the last criterion and first covariate trials; how many depends on numerous characteristics which cannot be known until the experiment has been run. If too many trials are introduced as buffer trials, one may increase the subjects' abilities markedly beyond what it had been when the criterion task was performed, a condition not acceptable when the covariate is to be used for bias reduction purposes.

This "Catch 22" only furthers the argument that the use of covariates to reduce subject bias is fraught with problems that reduce their effectiveness for subject bias reduction.

INTERPRETING THE MEASURES

When some form of the task itself is used as the covariate, an empirical demonstration of "validity" will probably not be necessary. The test has an inferred validity. What does it mean then if it turns out that such a test correlates poorly with the experimental performance scores from which ability is to be partialled? It does not mean that the validity of the covariate derived from the task itself is invalid. It merely means that for the specific set of data from which ability is being partialled, only a small amount of whatever the covariate measures -- subject ability -- is actually present. The square of the correlation is the proportion of total variance accounted for by individual differences in that single case.

RELIABILITY

The low correlation in the example above might be due to the poor reliability of the data employed. If either set of scores is unreliable, the usefulness of a covariate is impaired. The corrective measures for this possibility must be considered carefully; however, solutions to problems of unreliable measures should properly be resolved before the experiment is begun, not after it is over.

If the covariate task yields unreliable measures, then it is appropriate to use multiple measures to obtain a reliable average series of scores for ranking subjects according to ability. On the other hand, it may not be appropriate to use the average of performance measures made over several trials in the experiment.

One cannot toy with the performance scores from which subject ability is to be isolated. If the investigator intends to use only one set of measures from each subject to evaluate the experimental training or transfer effects, however unreliable that performance may be, those are the scores from which the covariate must be partialled. If one is dealing with real-world problems, one may want to predict or understand performance on the single trial (whereby a poor performance may negate the possibility of ever having a second trial, e.g., carrier landing).

Nor should the correlation between the covariate and performance be adjusted statistically to correct for poor reliability when that value is to be used to partial out ability. The new value would represent what the validity of the covariate ought to be, rather than what it was at the time the experimental data were being taken. Of course, if the performance scores have low reliability, it will not be possible to remove any effective amount of subject variability.

EVALUATION

In the two holistic transfer of training experiments by Simon and Roscoe (1981) and Westra (1982), the covariate method was used to partial out subject bias. The appropriateness of that approach must now be questioned.

Simon and Roscoe (1981) used as their covariate the median score of three trials taken at the beginning of the experiment on a common experimental condition at the center of the experimental space. Since the experiment was a quasi-transfer study with both training and transfer performed in the simulator, the validity of the covariate task was axiomatic.

The covariate they used, however, does not meet the requirement regarding the level of subject skill. Considering the unstable nature of performance when an operator is starting to learn a complex task, the ordering of subjects on their ability when tested early in the training program need not correlate well with that after learning has taken place. In addition to the exploration and testing of varied techniques that an operator often attempts early in a training program, one might also expect that some factors affecting skill at the end of the training phase may not even be operating at the beginning. The covariate used in this study may not have provided a relative measure of subject skill as it was at the time when the critical training and transfer trials were being performed.

In the Westra (1982) experiment, an Atari Air Combat Maneuvering game was used as the covariate. The validity of this task was never seriously investigated. It is certainly a much simpler task than that of carrier landing and there is no way in which to equate performance on it with skill level at a particular time during carrier landing training. After the fact, correlations between .4Ø and .55 were found between the Atari performance scores and the transfer scores of the different subjects, presumably after whatever transfer from the different training conditions that did occur had dissipated. How this was determined was never explained.

For both studies, the correlations from which the validity of the covariates were inferred were all below .55. Even at that level, the correlation would barely account for 2Ø% of the subject bias. Any of the inadequacies discussed earlier may have been responsible for the correlations not being higher. Even if a high correlation had been found between the covariate and a validation score (during some preexperimental development phase), accepting it merely because it was numerically high overlooks the possibility that the correlation might be due to the presence of a third variable which affected both measures which were themselves unrelated. That variable may not be present under operational conditions.

SUMMARY

Developing a covariate that suitably measures subject ability on a particular task at a particular time is difficult and expensive to achieve. Historically the validity coefficients obtained have been too low to be effective for removing subject bias. If a high correlation were ever achieved, it still would be necessary to reevaluate the validity coefficients when the situation context changes in another experiment or for a different subject population.

APPENDIX E

PROBABLE LOCATION OF THE TRUE SUBJECT BIAS EFFECTS WHEN LOCATIONS OF COVARIATE BIAS EFFECTS ARE KNOWN

The results of the Monte Carlo simulation for Technique 4 described in Section IV of the text are presented in Table E-1 below. The following examples illustrate how these tables are interpreted:

1. Table E-1.A, K = 31, r = .30, Row 1, Column 1

There is a .09 probability that the largest positive true bias effect will actually be confounded with the same factor effect as the largest positive covariate bias effect.

2. Table E-1.A, K = 31, r = .65, Row 1, Column 4

There is a .60 probability that the largest positive true bias effect will be confounded with one of the four factor effects that are confounded with the *four largest positive* covariate bias effects. Which one is not known.

3. Table E-1.A, K = 31, r = .707, Row 2, Column 8

There is a .64 probability that the two largest positive true bias effects will be confounded with two of the eight factor effects that are confounded with the eight largest positive covariate bias effects. Which two are not known.

4. Table E-1.B, K = 63, r = .45, Row 2, Column 8

There is a .17 probability that the two largest positive true bias effects will be confounded with two of the eight factor effects that are confounded with the eight largest positive covariate bias effects. This probability drops to .06 if the largest four covariate bias effects are used (Row 2, Column 4).

5. Table E-1.C, K = 127, r = .30, Row 1, Column 8

There is a .22 chance that the largest positive true bias effect will be confounded with one of the eight factor effects confounded with the eight largest positive covariate bias effects. That does not provide much effective use of the covariates to interpret confounded factor effects unless the number of covariate effects considered are increased.

Note that there is not always the same number of rows in every subtable. This is because no row is shown when all of its probabilities are less than .1Ø (when only eight or fewer covariate effects are being considered).

TABLE E-1. PROBABILITIES THAT TRUE SUBJEJCT BIAS EFFECTS ARE CONFOUNDED WITH FACTOR EFFECTS INDICATED BY THE LOCATION OF COVARIATE BIAS EFFECTS.*

		NUMBER OF LARGEST POSITIVE COVARIATE BIAS EFFECTS							
NUMBER OF LARGEST POSITIVE TRUE BIAS EFFECTS	RANK	1	2	3	4	5	6	7	8
A									
K=31; r=.30									
	1	.09	.17	.24	.30	.35	.40	.45	.49
	2		.01	.04	.07	.10	.13	.17	.21
K=31; r=.45									
	1	.14	.25	.33	.40	.47	.53	.58	.62
	2		.03	.07	.12	.17	.22	.27	.33
	3			.01	.02	.05	.07	.11	.14
K=31; r=.65									
	1	.27	.41	.52	.60	.67	.72	.77	.81
	2		.08	.17	.26	.35	.43	.50	.57
	3			.03	.08	.15	.21	.28	.35
	4				.01	.14	.08	.13	.18
K=31; r=.707									
	1	.31	.48	.59	.67	.73	.78	.82	.85
	2		.12	.23	.34	.43	.51	.58	.64
	3			.05	.12	.19	.27	.35	.42
	4				.02	.06	.12	.18	.25
	5					.01	.04	.07	.12

*Rows are discontinued when all probabilities are less than .10.

Table E-1 (Continued)

NUMBER OF LARGEST POSITIVE TRUE BIAS EFFECTS

		NUMBER OF LARGEST POSITIVE COVARIATE BIAS EFFECTS							
	RANK	1	2	3	4	5	6	7	8
B									
K=63; r=.30									
	1	.06	.11	.16	.20	.23	.26	.29	.32
K=63; r=.45									
	1	.11	.19	.25	.30	.35	.39	.43	.47
	2		.02	.04	.06	.10	.12	.14	.17
K=63; r=.65									
	1	.22	.34	.43	.50	.56	.61	.65	.69
	2		.05	.12	.18	.24	.30	.35	.39
	3			.02	.04	.08	.11	.14	.18
K=63; r=.707									
	1	.26	.40	.50	.57	.63	.68	.72	.75
	2		.08	.16	.23	.30	.37	.42	.48
	3			.02	.06	.11	.16	.21	.25
	4				.01	.03	.05	.08	.11
C									
K=127; r=.30									
	1	.04	.07	.10	.13	.15	.17	.19	.22
K=127; r=.65									
	1	.18	.28	.36	.42	.47	.50	.54	.57
	2		.03	.07	.12	.15	.19	.23	.26
	3			.01	.02	.04	.06	.08	.10
K=127; r=.707									
	1	.22	.34	.42	.49	.54	.59	.62	.66
	2		.05	.11	.17	.21	.26	.31	.34
	3			.01	.04	.06	.09	.12	.15

NOTATIONS AND TERMINOLOGY

Symbols

B1 Measure of skewedness = $(\Sigma x^3)^2/N^2s^6$

B2 Measure of kurtosis = $\Sigma x^4/Ns^6$

C Number of different experimental conditions in the experiment. In the basic design of holistic experiments, $C = 2^{k-p}$ conditions.

K Number of estimable effects in the experiment. $K = 2^{k-p} - 1 = C-1$.

k Number of factors to be investigated in an experiment. Maximum possible in Resolution IV design, $k = C/2$.

N Total number of independent observations or total number of subjects in an experiment. $N = nC = n2^{k-p}$.

"N" Used to differentiate the use of the N in published tables in order statistics from the use of N above, where "N" in published tables is equivalent to either C or K, depending on whether it is conditions or effects the investigator is relating to chance values.

n Number of subjects in every cell of the experiment. Equals number of replications.

$\underline{N}$[Ø,1] The underlined $\underline{N}$ is used to represent a normally distributed population, and in this example, with a mean of zero and a variance of one. Other values might be substituted for the mean and variance.

P Reduction in population variance due to percent bitruncation. P = [1 - New variance due to bitruncation divided by Old variance].

p Indicates what fraction of a full factorial for k factors the fractional factorial, 2^{k-p}, is. The fraction equals to $1/(2^p)$ of 2^k conditions.

R_i Rank position i in any ordered set of values.

r_{xy} Correlation coefficient between performance and covariate scores.

r_{yt}	Validity coefficient; correlation coefficient between covariate and "true" ability scores.
s	Standard deviation $s = \sqrt{V}$
T, t	Theoretical "true" ability measurements.
V	Variance. $V = s^2$
X, x	Performance scores
Y, y	Covariate scores
•	Symbol for multiplication process in equations.

Designs

Between-subjects

Different subject tested on each experimental condition.

Within-subject

Same subjects tested on all experimental conditions or all conditions within blocks of the full experiment.

Fully saturated

When the number of independent main and interaction effects being studied is one less than the number of experimental conditions (i.e., $K = C - 1$).

Fractional factorial

A selected portion of a factorial design. In the context of this report, fractional factorials are in the form of 2^{k-p}.

Holistic

A design that embodies the philosophy and strategy of a holistic approach: i.e., evolutionary and economical; accounts for all potentially important sources of variance whether of interest or not; provides tests of assumptions and lack of fit; emphasizes elimination of irrelevant effects.

Resolution IV

Experimental design in which all main effects are isolated from each other and from all two-factors interaction effects, while aliased two-factor interaction effects occur in independent strings.

Experiments

Few-factor

Looks at fewer than five factors in an experiment generally using a complete factorial design.

Many-factor

Looks at at least seven but generally more factors in a single experiment using economical designs.

Statistics

Sample subject (bias) effect

The mean difference between abilities of groups of subjects assigned at random to the two levels of a factor in a 2^{k-p} design. Effect is confounded with a factor effect.

Expected subject (bias) effect

Same as subject bias effect except that it is the average (mean) over many samples.

z-score (of bias) or z-value

Standardized form of the bias at any rank.

Population variance and population standard deviation

Variance of abilities to perform a designated task for a theoretical subject population, estimated from the scores obtained in a finite sample. Standard deviation equals square root of variance.

Subject sample variance

Used to estimate population variance. Square root of variance, or standard deviation, can be estimated using selected values in a normal-order plot (Zahn, 1975).

Bias variance

Variance of all possible bias effects (mean differences between the two groups of randomly assigned subjects for each source of variance with one degree of freedom). Is equal to the population variance times (4/N) where N is equal to the total number of subjects. Standard deviation of this value is often referred to as the standard error of the mean difference.

Standard deviation of bias

Standard deviation of an individual bias effect at rank i is equal to the standard deviation of the z-score for rank i of k cases times the population standard deviation times the square root of (4/N).

Analysis and Interpretation

Normal Order Plot

A plot of data on normal probability paper used to judge whether or not observed effects are likely to have occurred by chance.

LIST OF EQUATIONS

1) $$F = \frac{\text{B/Cell Var. (i.e., Factor Var. + Subj. Var. b/cells + Error Var.)}}{\text{W/Cell Var. (i.e., Subj. Var. w/cells + Error Var.)}}$$

where F = F ratio; B/ and W/ = Between and Within; and Var. = Variance

2) Between Cell Var. (i.e., Factor Var. + Subj. Var. b/Cells + Error Var.)

3) $$\begin{array}{l}\text{Expected z-score}\\ \text{of bias effect}\\ \text{for rank i of}\\ \text{K cases}\end{array} = \begin{array}{l}\text{Expected bias for}\\ \text{rank i of K cases}\\ \text{from a population}\\ \text{N(0,1)}\end{array} \div \left[\begin{array}{l}\text{Population}\\ \text{Standard}\\ \text{Deviation}\end{array} \cdot \sqrt{\frac{4}{N}}\right]$$

where K cases are the number of independent effects being seriously examined and N is the total number of subjects used.

4) $$\begin{array}{l}\text{Expected bias}\\ \text{for rank i of}\\ \text{K cases}\end{array} = \begin{array}{l}\text{Expected z-score}\\ \text{for rank i of K}\\ \text{cases from popu-}\\ \text{lation }\underline{\text{N}}\text{(0,1)}\end{array} \cdot \left[\begin{array}{l}\text{Population}\\ \text{Standard}\\ \text{Deviation}\end{array} \cdot \sqrt{\frac{4}{N}}\right]$$

5) $$\begin{array}{l}\text{Standard devia-}\\ \text{tion of bias for}\\ \text{rank i of K cases}\end{array} = \begin{array}{l}\text{Standard deviation}\\ \text{of the z-score for}\\ \text{rank i of K cases}\end{array} \cdot \left[\begin{array}{l}\text{Population}\\ \text{Standard}\\ \text{Deviation}\end{array} \cdot \sqrt{\frac{4}{N}}\right]$$

6) $$\begin{array}{l}\text{Proportion of}\\ \text{variance for}\\ \text{rank i of K}\\ \text{cases}\end{array} = \frac{(\text{z-score at rank i})^2}{\text{Sum of (z-scores)}^2}$$

List of Equations (Continued)

7) $$\text{Performance score with covariate effect removed} = \text{Performance Score (X)} - \left[r_{yx} \left(\frac{sX}{sY} \right) \bullet \text{Covariate Score (Y)} \right]$$

where s = standard deviation and r_{yx} is correlation between X and Y

8) $$\text{Adjusted Bias (covariate effectiveness)} = \text{Original Bias} \bullet \sqrt{1 - r_{yt}^2}$$

where r_{yt} is the correlation between the covariate scores (Y) and the "true" measures (T) of subject ability.

9) $$\text{Percent bias reduction} = \left[1 - \frac{\text{Adjusted Bias}}{\text{Original Bias}} \right] \bullet 100$$

10) $$\text{Adjusted bias (bitruncated)} = \text{Original bias} \bullet \sqrt{1 - \left(r_{yt}^2 \bullet P \right)}$$

where P is the proportion of variance reduction due to bitruncation, i.e., P = (1 — New variance due to bitruncation/Old variance).

11) $$\text{Adjusted Bias (replication)} = \text{Original Bias} \bullet \sqrt{\frac{\text{N original}}{[\text{N orig.} \bullet \text{No. Replicates}]}}$$

12) $$\text{Percent bias reduction (replication)} = \left[1 - \sqrt{1/n} \right] \bullet 100$$

where n is the number of times the original design was replicated. This n also equals the number of subjects per experimental condition.